TITLE OF THE INVENTION 2318-380A

CHROMOSOME 17p-LINKED PROSTRATE CANCER SUSCEPTIBILITY GENE AND A PARALOG AND ORTHOLOGOUS GENES

CROSS REFERENCE TO RELATED APPLICATION

This application is a divisional of Serial No. 09/564,805, filed 5 May 2000, which is a continuation-in-part of Serial No. 09/434,382, filed 5 November 1999 and is related to U.S. provisional patent application Serial No. 60/107,468, filed 6 November 1998, and priority is claimed thereto under 35 U.S.C. §119(e). Each of these applications is incorporated herein by reference.

This application was made with Government support under Grant Nos. CA62154 and CA64477 from the National Institutes of Health. The United States Government has certain rights in this invention.

BACKGROUND OF THE INVENTION

The publications and other materials used herein to illuminate the background of the invention, and in particular, cases to provide additional details respecting the practice, are incorporated herein by reference, and for convenience, are referenced by author and date in the following text and respectively grouped in the appended List of References.

The genetics of cancer is complicated, involving the function of three loosely defined classes of genes: (1) dominant, positive regulators of the transformed state (oncogenes); (2) recessive, negative regulators of the transformed state (tumor suppressor genes); and (3) genes that modify risk without playing a direct role in the biology of transformed cells (risk modifiers).

Specific germline alleles of certain oncogenes and tumor suppressor genes are causally associated with predisposition to cancer. This set of genes is referred to as tumor predisposition genes. Some of the tumor predisposition genes which have been cloned and characterized influence susceptibility to: 1) Retinoblastoma (RB1); 2) Wilms' tumor (WT1); 3) Li-Fraumeni (TP53); 4) Familial adenomatous polyposis (APC); 5) Neurofibromatosis type 1 (NF1); 6) Neurofibromatosis type 2 (NF2); 7) von Hippel-Lindau syndrome (VHL); 8) Multiple endocrine neoplasia type 2A (MEN2A); 9) Melanoma (CDKN2 and CDK4); 10) Breast and ovarian cancer (BRCA1 and BRCA2); 11) Cowden disease (MMAC1); 12) Multiple endocrine neoplasia (MEN1); 13) Nevoid basal cell carcinoma syndrome (PTC); 14) Tuberous sclerosis 2

(TSC2); 15) Xeroderma pigmentosum (genes involved in nucleotide excision repair); 16) Hereditary nonpolyposis colorectal cancer (genes involved in mismatch repair).

Specific germline alleles of certain risk modifier genes are also associated with predisposition to cancer, but the increased risk is sometimes only clearly expressed when it is combined with certain environmental, dietary, or other factors. Alcohol dehydrogenase (ADH) oxidizes ethanol to acetaldehyde, a chemical which is both mutagenic and carcinogenic in lab animals. The enzyme encoded by the $ADH3^1$ allele oxidizes ethanol relatively quickly, whereas the enzyme encoded by the $ADH3^2$ allele oxidizes ethanol more slowly. $ADH3^1$ homozygotes presumably have a high capacity for synthesis of acetaldehyde; those who also drink heavily are at increased risk for oral cavity, esophageal, and (in women) breast cancer relative to $ADH3^2$ homozygotes who drink equally heavily (Harty et al., 1997; Hori et al., 1997; Shields, 1997). The acetyltransferases encoded by N-acetyltransferase 1 (NAT1) and N-acetyltransferase 2 (NAT2) catalyze the acetylation of numerous xenobiotics including the aromatic amine carcinogens derived from smoking tobacco products. Individuals who are homozygous for slow acetylating forms of NAT2 who are also heavy smokers are at greater risk for lung, bladder, and (in females) breast cancer than individuals who smoke equally heavily but are homozygous for fast acetylating forms of NAT2 (Shields, 1997; Bouchardy et al., 1998).

The risk of hormone related cancers such as breast and prostate cancer may be modulated by allelic variants in enzymes that play a role in estrogen or androgen metabolism, or variants in proteins that mediate the biological effects of estrogens or androgens. A polymorphic CAG repeat in the human androgen receptor gene encodes a polymorphic polyglutamine tract near the amino-terminus of the protein. The length of the polyglutamine tract is inversely correlated with the transcriptional activation activity of the androgen receptor and thus one aspect of the biological response to androgens. Men whose androgen receptor contains a relatively short polyglutamine tract are at higher risk for prostate cancer, especially high stage/high histologic grade prostate cancer, than men whose androgen receptor contains a relatively long polyglutamine tract (Giovannucci et al., 1997).

Prostate cancer is the most common cancer in men in many western countries, and the second leading cause of cancer deaths in men. It accounts for more than 40,000 deaths in the US annually. The number of deaths is likely to continue rising over the next 10 to 15 years. In the US, prostate cancer is estimated to cost $1.5 billion per year in direct medical expenses. In addition to the burden of suffering, it is a major public-health issue. Numerous studies have

provided evidence for familial clustering of prostate cancer, indicating that family history is a major risk factor for this disease (Cannon et al., 1982; Steinberg et al., 1990; Carter et al, 1993).

Prostate cancer has long been recognized to be, in part, a familial disease with a genetic component (Woolf, 1960a; Cannon et al., 1982; Carter et al., 1992). Numerous investigators have examined the evidence for genetic inheritance and concluded that the data are most consistent with dominant inheritance for a major susceptibility locus or loci. Woolf (1960b), described a relative risk of 3.0 of developing prostate cancer among first-degree relatives of prostate cancer cases in Utah using death certificate data. Relative risks ranging from 3 to 11 for first-degree relatives of prostate cancer cases have been reported (Cannon et al., 1982; Woolf, 1960b; Fincham et al., 1990; Meikle et al., 1985; Krain, 1974; Morganti et al., 1956; Goldgar et al., 1994). Carter et al. (1992) performed segregation analysis on families ascertained through a single prostate cancer proband. The analysis suggested Mendelian inheritance in a subset of families through autosomal dominant inheritance of a rare (q=0.003), high-risk allele with estimated cumulative risk of prostate cancer for carriers of 88% by age 85. Inherited prostate cancer susceptibility accounted for a significant proportion of early-onset disease, and overall was responsible for 9% of prostate occurrence by age 85. Recent results demonstrate that at least four loci exist which convey susceptibility to prostate cancer as well as other cancers. These loci are *HPC1* on chromosome 1q24, (Smith et al., 1996), *HPCX* on chromosome Xq27-28 (Xu et al., 1998), *PCAP* at 1q42 (Berthon et al., 1998) and *CAPB* at 1p36 (Gibbs et al., 1999a). All four suggestions of linkage for prostate cancer predisposition were the result of hints arising from genome-wide searches. Although only the *HPC1* linkage has so far been confirmed (Cooney et al., 1997; Neuhausen et al., 1999; Xu and ICPCG, 2000), it is becoming clear that a large number of genes contribute to familial prostate cancer. It also seems clear, both from published hereditary prostate cancer linkage studies and from genotyping of our family resource at the above mentioned loci, that no single predisposition locus mapped to date is by itself responsible for a large portion of familial prostate cancer (Neuhausen et al., 1999; Eeles et al., 1998; Gibbs et al., 1999b; Lange et al., 1999; Berry et al., 2000; Suarez et al., 2000; Goode et al., 2000).

A comparison to the cloning of, and risk profile attributed to, breast cancer susceptibility genes provides an instructive example. The profusion of proposed prostate loci, coupled with minimal confirmation or refined localization following initial publication of these linkages, contrasts sharply with studies of the breast and ovarian cancer susceptibility genes *BRCA1* and *BRCA2*. Linkage to *BRCA1* was first published in 1990 (Hall et al., 1990); groups competing to

identify this gene moved swiftly from confirmatory studies through efforts to refine the localization to the gene identification in 1994 (Miki et al., 1994). With expanded genomics resources, the time from linkage (Wooster et al., 1994) to complete cloning (Wooster et al., 1995; Tavtigian et al., 1996) of *BRCA2* was only slightly more than 1 year. Ongoing mutation screening and careful modeling of age specific and familial risks indicate that these two genes account for virtually all extended breast and ovarian cancer families (Antoniou et al., 2000) and the majority of breast cancer families with more than five cases, especially those that include an early-onset component (Ford et al., 1998).

Even so, a fraction of familial breast cancer risk is manifest in smaller family clusters with average age at diagnosis. While *BRCA1* and *BRCA2* only account for a portion of this component of breast cancer risk (Peto et al., 1999), there are no published and confirmed linkages based on these types of pedigrees to date. Standard genetic analysis appears to be limited by the problems of low penetrance and genetic complexity. It is possible that analysis of genetic predisposition in families with excess prostate cancer also reflects these issues. As absence of distinction by age at diagnosis/onset would also be consistent with the influence of multiple susceptibility genes harboring only moderate risk sequence variants, one might therefore ask what relative contribution low frequency high risk variants analogous to mutations in *BRCA1/2*, versus higher frequency, moderate risk sequence variants, make to the population risk of prostate cancer.

Indeed, evidence that moderate risk sequence variants in a number of specific genes contribute to prostate cancer susceptibility is beginning to accumulate. For example, a polymorphic CAG repeat within the androgen receptor open reading frame encoding a variable length polyglutamine tract shows an inverse relationship between repeat length and the transcriptional transactivation activity of the receptor (Chamberlain et al., 1994; Kazemi-Esfarjani et al., 1995). Accordingly, a series of studies show an association between shorter androgen receptor CAG repeat length and prostate cancer risk (Giovannucci et al., 1997; Stanford et al., 1997), although it is not entirely clear whether the association is with diagnosis of prostate cancer or severity of disease (Bratt et al., 1999). Second, a number of missense variants have been observed in the steroid 5α-reductase type II gene (*SRD5A2*), responsible for conversion of testosterone to the more active androgen dihydrotestosterone in the prostate (Makridakis et al., 1997). One of these variants, Ala 49 Thr, has been reported to increase the catalytic activity of the enzyme, and is associated with increased risk of advanced prostate cancer (Makridakis et al., 1999; Jaffe et al., 2000). Finally, several groups have reported an

excess of prostate cancer in large *BRCA2* pedigrees (Sigurdsson et al., 1997; Breast Cancer Linkage Consortium, 1999), though the relative risk that these mutations confer for prostate cancer is considerably lower than for breast cancer. Further, these effects may be variant specific as association has not been confirmed among men who carry the Ashkenazi BRCA2 founder mutation 6174delT (Wilkens et al., 1999; Nastiuk et al., 1999; Hubert et al., 1999). If these and similar sequence variants play a role in a significant fraction of prostate cancer, then models of the genetic component of familial prostate cancer may need to incorporate both linkage evidence for major susceptibility loci and association evidence for moderate risk sequence variants.

The Utah population provides a unique resource for examining the genetic basis of disease. Extended high risk pedigrees containing many cases can be ascertained as units instead of by expansion from individual probands. While these pedigrees are an extremely powerful resource for linkage studies, they also allow analysis of segregation of moderate risk sequence variants through multiple generations of both cases and their unaffected relatives.

Detection of genetic linkage for prostate cancer susceptibility to a defined segment of a chromosome requires that DNA sequence variants within that chromosomal segment confer the cancer susceptibility. This is usually taken to mean that the causal sequence variant(s) will either alter the expression of one or more linked genes or will alter the function of one of the linked genes. However, detection of the genetic linkage does not necessarily provide evidence for what class of gene (i.e. tumor suppressor, oncogene, or risk modifier) is affected by the causal sequence variant(s).

Most strategies for proceeding from genetic linkage of prostate cancer susceptibility to chromosome 17p to identification of the 17p-linked prostate cancer predisposing gene (HPC2) require precise genetic localization studies to define a discrete segment of the chromosome within which the causal sequence variant(s) must map. Gene identification projects based on precise genetic localization are called positional cloning projects. The general strategy in positional cloning is to find all of the genes located within the genetically defined interval, identify sequence variants in and around those genes, and then determine which of those sequence variants either alter the expression or the function of one (or more) of the associated genes. Segregation of such sequence variants with the disease in the linked kindreds must also be demonstrated. We have executed a positional cloning project in the HPC2 region of chromosome 17p and found a gene, herein named HPC2, germline mutations which predisposes individuals to prostate cancer.

SUMMARY OF THE INVENTION

The present invention relates generally to the field of human genetics. Specifically, the present invention relates to methods and materials used to isolate and detect a human prostate cancer predisposing gene (HPC2), some alleles of which cause susceptibility to cancer, in particular prostate cancer. More specifically, the present invention relates to germline mutations in the HPC2 gene and their use in the diagnosis of predisposition to prostate cancer. The invention also relates to presymptomatic therapy of individuals who carry deleterious alleles of the HPC2 gene. The invention further relates to somatic mutations in the HPC2 gene in human prostate cancer and their use in the diagnosis and prognosis of human prostate cancer. Additionally, the invention relates to somatic mutations in the HPC2 gene in other human cancers and their use in the diagnosis and prognosis of human cancers. The invention also relates to the therapy of human cancers which have a mutation in the HPC2 gene, including gene therapy, protein replacement therapy, protein mimetics, and inhibitors. The invention further relates to the screening of drugs for cancer therapy. The invention also relates to the screening of the HPC2 gene for mutations, which are useful for diagnosing the predisposition to prostate cancer. The HPC2 gene is useful as a marker for the HPC2 locus and as a marker for prostate cancer. Finally, a paralog of HPC2 as well as orthologs of HPC2 in mouse, chimpanzee and gorilla have been isolated and characterized.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 is a multipoint linkage analysis of 4 kindreds that show suggestive evidence for linkage to the HPC2 prostate cancer susceptibility locus relative to chromosome 17p markers.

Figures 2A-B are diagrams showing the order of genetic markers and recombinant boundaries neighboring HPC2, a schematic map of BACs spanning the HPC2 region, a schematic map of transcription units within the HPC2 region, and two diagrams of the HPC2 transcription unit showing the locations of the exons of HPC2 relative to the BAC to which it maps and relative to each other. The individual exons are numbered.

Figure 3 shows recombinant, physical and transcript maps centered at the human *ELAC2* locus on chromosome 17p. The top portion shows genetic markers and recombinants. Microsatellite markers developed at Myriad Genetics, Inc. are given as 17-MYR####. Nested within the arrows that represent meiotic recombinants are the pedigree in which the recombinant occurred and, in parentheses, the number of cases who carry the haplotype on which the recombinant occurred. The second portion of the figure shows a BAC contig tiling path across

this interval. The T7 end of each BAC is denoted with an arrowhead. The third portion of the figure shows transcription units identified in the interval. The bottom portion of the figure is an expanded view of a 40 kb segment at the SP6 end of BAC 31k12 showing the relative positions of two exons of the gene 04CG09 and all of the coding exons of *ELAC2*.

Figure 4 is an alignment of the sequence of exon 1 of the human HPC2 gene with exon 1 of the mouse HPC2 gene. The figure also shows an alignment of the peptide sequence encoded by exon 1 of the human HPC2 gene with the peptide sequence encoded by exon 1 of the mouse HPC2 gene. The human DNA sequence is SEQ ID NO:210; the human amino acid sequence is SEQ ID NO:211; the mouse DNA sequence is SEQ ID NO:212 and the mouse amino acid sequence is SEQ ID NO:213.

Figures 5A-B show kindreds 4102 and 4289. The pedigrees have been genotyped over a 20 cM interval extending from D17S786 to D17S805. Haplotypes are represented by the bars; the dark gray haplotype segregating in each pedigree is the mutation bearing chromosome. The relative position of *ELAC2* is denoted by * (white on black or white on gray). Figure 5A shows kindred 4102. The dark bar denotes the 1641 insG bearing haplotype. Individuals 061 and 107 carry part of the frameshift haplotype, but neither carries the frameshift due to recombination events. There are no data to distinguish which of the founders, individuals 050 and 051, carried the frameshift. The second shared haplotype in kindred 4102 is denoted by a light gray bar. Again, there are no data to distinguish which of the founders, individuals 005 and 006, carried the haplotype. Figure 5B shows kindred 4289. Individuals 064, 066, 067, 068 and 072 share a recombinant chromosome that carries the His 781 missense change.

Figures 6A-B are a multiple protein alignment of ELAC1/2 family members. ELAC2 family members were selected from human (HSA), mouse (MMU), *C. elegans* (CEL), *A. thaliana* (ATH) and *S. cerevisiae* (SCE). The *A. thaliana* genome encodes more than one family member; gi6850339 was selected because it aligned with fewer gaps. ELAC1 family members were selected from human, *E. coli* (Es_c), the blue-green algae *Synechocystis* (Syn) and the archaebacterium *Methanobacterium thermoautotrophicum* (Me_t). Alignments were based on BLASTp searches and then optimized by inspection. The positions of Ser 217, Ala 541 and Arg 781 in human ELAC2 are marked by down arrows. The sequences shown in Figures 6A-B are: human ELAC2 is SEQ ID NO:2, mouse Elac2 is SEQ ID NO:222, *C. elegans* CE16965 is SEQ ID NO:227, *A. thaliana* gi 6850339 is SEQ ID NO:228, *S. cerevisiae* YKR079C is SEQ ID NO:229, human ELAC1 is SEQ ID NO:220, *E. coli* elaC is SEQ ID NO:230, *Synechocystis*

gi2500943 is SEQ ID NO:231, and *Methanobacterium thermoautotrophicum* gi2622965 is SEQ ID NO:232.

Figure 7 shows recessive genotype frequencies by birth cohort.

Figure 8 shows the results of association tests.

Figure 9 shows a multiple protein alignment demonstrating conservation of sequence elements between ELAC2, PSO2 and CPSF73 families. The alignments shown for segments of PSO2 and CPSF73 family members were taken from more extensive alignments that contain family members from a larger set of species, analogous to the ELAC1/2 alignments of Figures 6A-B. The seven His or Cys residues that are conserved across two or more of the gene families are marked by down arrows. The position of Ala 541 in human ELAC2 is also marked by a down arrow. The sequences shown in Figure 9 are partial sequences of the following: human CPSF73 is SEQ ID NO:233, *A. thaliana* gi6751699 is SEQ ID NO:234, *S. cerevisiae* YSH1 is SEQ ID NO:235, *Synechocystis* gi2496795 is SEQ ID NO:236, *Methanobacterium thermoautotrophicum* gi2622312 is SEQ ID NO:237, human ha3611 is SEQ ID NO:238, *A. thaliana* gi2979557 is SEQ ID NO:239, *S. cerevisiae* PSO2 is SEQ ID NO:240, human ELAC2 is SEQ ID NO:2, *A. thaliana* gi6850339 is SEQ ID NO:228, and *S. cerevisiae* YKR079C is SEQ ID NO:229.

Figure 10 shows a similarity comparison among the ELAC2 family members aligned in Figures 6A-B.

Figures 11A-D shows an analysis of ELAC1 expression in human tissues. Figures 11A-B show Multiple Tissue Northern (MTN) filters (Clontech) probed with the human ELAC1 ORF. Note that a 3 kb ELAC1 transcript is detected in all tissues. Figures 11C-D show the same filters probed with human β-actin as a loading control.

Figure 12 shows a multiple protein alignment demonstrating similarity between an N-terminal segment of the ELAC2 family members and the sequence context of the histidine motif shared by ELAC1 and ELAC2 family members. Species abbreviations are as in Figures 6A-B. The sequences shown in Figure 12 are partial sequences of the following: human ELAC2 is SEQ ID NO:2, mouse Elac2 is SEQ ID NO:222, *C. elegans* CE16965 is SEQ ID NO:227, *A. thaliana* gi6850339 is SEQ ID NO:228, *S. cerevisiae* YKR079C is SEQ ID NO:229, human ELAC1 is SEQ ID NO:220, *E. coli* elaC is SEQ ID NO:230, *Synechocystis* gi2500943 is SEQ ID NO:231, and *Methanobacterium thermoautotrophicum* gi2622965 is SEQ ID NO:232.

Figure 13 shows the relationship between ELAC1/2, PSO2 and CPSF73 gene family members. The tree is a distance-based depiction of pairwise sequence similarities determined

from a manual alignment of the ~67 amino acids immediately surrounding the histidine motif. ClustalX (Thompson et al., 1997) was used to calculate the percent divergence of each sequence on a pairwise basis and neighbor joining (Saitou and Nei, 1987) was applied to the resulting distance matrix. The treefile produced from ClustalX was visualized using TreeView (Page, 1996) and further edited in a graphics program for aesthetics. The scale bar indicates amino acid substitutions per residue.

BRIEF DESCRIPTION OF THE TABLES

Table 1 is a compilation of 2-point LOD scores for markers in the HPC2 region.

Table 2A lists the family resource used to detect linkage of *HPC2* to chromosome 17p.

Table 2B lists two-point LOD scores using the Utah age-specific model.

Table 3 is a summary of resource genotyped for the association tests.

Table 4 is a list of the accession numbers of human EST sequences used to assemble a tentative, partial cDNA sequence of the human HPC2 gene.

Table 5 is a list of the primers used for obtaining 5' RACE products that contained the start codon and part of the 5' UTR of the human HPC2 gene, primers used to prepare a full length human HPC2 expression construct, and primers used to check the sequence of that construct.

Table 6 is a list of the accession numbers of mouse EST sequences used to assemble a tentative, partial cDNA sequence of the mouse HPC2 gene.

Table 7 is a list of the primers used for obtaining 5' RACE products that contained the start codon and part of the 5' UTR of the mouse HPC2 gene, primers used to prepare a full length mouse HPC2 expression construct, and primers used to check the sequence of that construct.

Table 8 is a list of the primers used to mutation screen the human HPC2 gene from genomic DNA.

Table 9 is a summary of germline sequence variants of the human HPC2 gene.

Table 10 is a list of the allele frequencies of *HPC2*.

SUMMARY OF SEQUENCE LISTING

SEQ ID NO:1 is the nucleotide sequence for the human HPC2 cDNA from the start codon through the stop codon.

SEQ ID NO:2 is the amino acid sequence for the human HPC2 protein.

SEQ ID NO:3 is the nucleotide sequence for the human HPC2 cDNA from 50 base pairs before the start codon through the end of the 3' UTR.

SEQ ID NO:4 to SEQ ID NO:27 are the sequences of exon 1 to exon 24 of the human HPC2 gene.

SEQ ID NO:28 is the genomic sequence of the human HPC2 gene.

SEQ ID NOs:29-190 are nucleotide sequences of primers used to identify the human and/or mouse HPC2 genes or to screen for mutations.

SEQ ID NOs:191-209 are nucleotide sequences of the HPC2 around and including various sequence variants.

SEQ ID NO:210 is the nucleotide sequence of human HPC2 exon 1 and SEQ ID NO:211 is the corresponding amino acid sequence as shown in Figure 4.

SEQ ID NO:212 is nucleotide sequence of mouse HPC2 exon 1 and SEQ ID NO:213 is the corresponding amino acid sequence as shown in Figure 4.

SEQ ID NO:214 is a histidine containing motif found in HPC2/ELA2 and ELAC1.

SEQ ID NO:215 is exon 1 of ELAC1.

SEQ ID NO:216 is exon 2 of ELAC1 plus surrounding genomic sequence.

SEQ ID NO:217 is exon 3 of ELAC1 plus surrounding genomic sequence.

SEQ ID NO:218 is exon 4 of ELAC1 plus surrounding genomic sequence.

SEQ ID NO:219 is the cDNA for ELAC1 and SEQ ID NO:220 is the amino acid sequence for ELAC1.

SEQ ID NO:221 is the cDNA for mouse ELAC2 and SEQ ID NO:222 is the amino acid sequence for mouse ELAC2.

SEQ ID NO:223 is the cDNA for chimpanzee ELAC2 and SEQ ID NO:224 is the amino acid sequence for chimpanzee ELAC2.

SEQ ID NO:225 is the cDNA for gorilla ELAC2 and SEQ ID NO:226 is the amino acid sequence for gorilla ELAC2.

SEQ ID NOs:227-229 are the amino acid sequences for ELAC2 family member proteins from *C. elegans*, *A. thaliana* and *S. cerevisiae* as shown in Figure 6A.

SEQ ID NOs:230-232 are the amino acid sequences for ELAC1 family member proteins from *E. coli*, *Synechocystis* and *Methanobacterium thermoautotrophicum* as shown in Figures 6A-B.

SEQ ID NOs:233-240 are amino acid sequences of proteins from CPSF73 and PSO2 families as shown in Figure 9. These are, respectively, human CPSF73, *A. thaliana* gi6751699,

S. cerevisiae YSH1, *Synechocystis* gi2496795, *Methanobacterium thermoautotrophicum* gi2622312, human ha3611, *A. thaliana* gi2979557 and *S. cerevisiae* PSO2. The sequences for the ELAC2 family of Figure 9 are SEQ ID NO:2 for human, SEQ ID NO:228 for *A. thaliana* (as for Figures 6A-B) and SEQ ID NO:229 for *S. cerevisiae* (as for Figures 6A-B). The sequence listing shows the complete sequences of these proteins whereas Figure 9 shows only portions of each sequence.

DETAILED DESCRIPTION OF THE INVENTION

The present invention provides an isolated polynucleotide comprising all, or a portion of the HPC2 locus or of a mutated HPC2 locus, preferably at least eight bases and not more than about 27 kb in length. Such polynucleotides may be antisense polynucleotides. The present invention also provides a recombinant construct comprising such an isolated polynucleotide, for example, a recombinant construct suitable for expression in a transformed host cell.

Also provided by the present invention are methods of detecting a polynucleotide comprising a portion of the HPC2 locus or its expression product in an analyte. Such methods may further comprise the step of amplifying the portion of the HPC2 locus, and may further include a step of providing a set of polynucleotides which are primers for amplification of said portion of the HPC2 locus. The method is useful for either diagnosis of the predisposition to cancer or the diagnosis or prognosis of cancer. The HPC2 gene is useful as a marker for the HPC2 locus and as a marker for prostate cancer.

The present invention also provides isolated antibodies, preferably monoclonal antibodies, which specifically bind to an isolated polypeptide comprised of at least five amino acid residues encoded by the HPC2 locus.

The present invention also provides kits for detecting in an analyte a polynucleotide comprising a portion of the HPC2 locus, the kits comprising a polynucleotide complementary to the portion of the HPC2 locus packaged in a suitable container, and instructions for its use.

The present invention further provides methods of preparing a polynucleotide comprising polymerizing nucleotides to yield a sequence comprised of at least eight consecutive nucleotides of the HPC2 locus; and methods of preparing a polypeptide comprising polymerizing amino acids to yield a sequence comprising at least five amino acids encoded within the HPC2 locus.

The present invention further provides methods of screening the HPC2 gene to identify mutations. Such methods may further comprise the step of amplifying a portion of the HPC2

locus, and may further include a step of providing a set of polynucleotides which are primers for amplification of said portion of the HPC2 locus. Such methods may also include a step of providing the complete set of short polynucleotides defined by the sequence of HPC2 or discrete subsets of that sequence, all single-base substitutions of that sequence or discrete subsets of that sequence, all 1-, 2-, 3-, or 4-base deletions of that sequence or discrete subsets of that sequence, and all 1-, 2-, 3-, or 4-base insertions in that sequence or discrete subsets of that sequence. The method is useful for identifying mutations for use in either diagnosis of the predisposition to cancer or the diagnosis or prognosis of cancer.

The present invention further provides methods of screening suspected HPC2 mutant alleles to identify mutations in the HPC2 gene.

In addition, the present invention provides methods to screen drugs for inhibition or restoration of HPC2 gene product function as an anticancer therapy.

The present invention also provides the means necessary for production of gene-based therapies directed at cancer cells. These therapeutic agents may take the form of polynucleotides comprising all or a portion of the HPC2 locus placed in appropriate vectors or delivered to target cells in more direct ways such that the function of the HPC2 protein is reconstituted. Therapeutic agents may also take the form of polypeptides based on either a portion of, or the entire protein sequence of HPC2. These may functionally replace the activity of HPC2 in vivo.

Finally, the present invention provides the sequence of a paralog of HPC2, herein called ELAC1, as well as the sequences of HPC2 orthologs from mouse, chimpanzee and gorilla. These orthologs are named ELAC2.

It is a discovery of the present invention that the HPC2 locus which predisposes individuals to prostate cancer, is a gene encoding an HPC2 protein, which has been found to be non-identical to publicly available protein or cDNA sequences. This gene is termed HPC2 herein. It is a discovery of the present invention that mutations in the HPC2 locus in the germline are indicative of a predisposition to prostate cancer. Finally, it is a discovery of the present invention that germline mutations in the HPC2 locus are also associated with prostate cancer and other types of cancer. The mutational events of the HPC2 locus can involve deletions, insertions and nucleotide substitutions within the coding sequence and the non-coding sequence.

Useful Diagnostic Techniques

According to the diagnostic and prognostic method of the present invention, alteration of the wild-type HPC2 locus is detected. In addition, the method can be performed by detecting the wild-type HPC2 locus and confirming the lack of a predisposition to cancer at the HPC2 locus. "Alteration of a wild-type gene" encompasses all forms of mutations including deletions, insertions and point mutations in the coding and noncoding regions. Deletions may be of the entire gene or of only a portion of the gene. Point mutations may result in stop codons, frameshift mutations or amino acid substitutions. Somatic mutations are those which occur only in certain tissues, e.g., in the tumor tissue, and are not inherited in the germline. Germline mutations can be found in any of a body's tissues and are inherited. If only a single allele is somatically mutated, an early neoplastic state is indicated. However, if both alleles are somatically mutated, then a late neoplastic state is indicated. The finding of HPC2 mutations thus provides both diagnostic and prognostic information. An HPC2 allele which is not deleted (e.g., found on the sister chromosome to a chromosome carrying an HPC2 deletion) can be screened for other mutations, such as insertions, small deletions, and point mutations. It is believed that many mutations found in tumor tissues will be those leading to decreased expression of the HPC2 gene product. However, mutations leading to non-functional gene products would also lead to a cancerous state. Point mutational events may occur in regulatory regions, such as in the promoter of the gene, leading to loss or diminution of expression of the mRNA. Point mutations may also abolish proper RNA processing, leading to reduction or loss of expression of the HPC2 gene product, expression of an altered HPC2 gene product, or to a decrease in mRNA stability or translation efficiency.

Useful diagnostic techniques include, but are not limited to fluorescent *in situ* hybridization (FISH), direct DNA sequencing, PFGE analysis, Southern blot analysis, single stranded conformation analysis (SSCA), RNase protection assay, allele-specific oligonucleotide (ASO), dot blot analysis, hybridization using nucleic acid modified with gold nanoparticles and PCR-SSCP, as discussed in detail further below. Also useful is the recently developed technique of DNA microchip technology.

Predisposition to cancers, such as prostate cancer, and the other cancers identified herein, can be ascertained by testing any tissue of a human for mutations of the HPC2 gene. For example, a person who has inherited a germline HPC2 mutation would be prone to develop cancers. This can be determined by testing DNA from any tissue of the person's body. Most

simply, blood can be drawn and DNA extracted from the cells of the blood. In addition, prenatal diagnosis can be accomplished by testing fetal cells, placental cells or amniotic cells for mutations of the HPC2 gene. Alteration of a wild-type HPC2 allele, whether, for example, by point mutation or deletion, can be detected by any of the means discussed herein.

There are several methods that can be used to detect DNA sequence variation. Direct DNA sequencing, either manual sequencing or automated fluorescent sequencing can detect sequence variation. For a gene as large as HPC2, manual sequencing is very labor-intensive, but under optimal conditions, mutations in the coding sequence of a gene are rarely missed. Another approach is the single-stranded conformation polymorphism assay (SSCA) (Orita *et al.*, 1989). This method does not detect all sequence changes, especially if the DNA fragment size is greater than 200 bp, but can be optimized to detect most DNA sequence variation. The reduced detection sensitivity is a disadvantage, but the increased throughput possible with SSCA makes it an attractive, viable alternative to direct sequencing for mutation detection on a research basis. The fragments which have shifted mobility on SSCA gels are then sequenced to determine the exact nature of the DNA sequence variation. Other approaches based on the detection of mismatches between the two complementary DNA strands include clamped denaturing gel electrophoresis (CDGE) (Sheffield *et al.*, 1991), heteroduplex analysis (HA) (White *et al.*, 1992) and chemical mismatch cleavage (CMC) (Grompe *et al.*, 1989). None of the methods described above will detect large deletions, duplications or insertions, nor will they detect a regulatory mutation which affects transcription or translation of the protein. Other methods which might detect these classes of mutations such as a protein truncation assay or the asymmetric assay, detect only specific types of mutations and would not detect missense mutations. A review of currently available methods of detecting DNA sequence variation can be found in a recent review by Grompe (1993). Once a mutation is known, an allele specific detection approach such as allele specific oligonucleotide (ASO) hybridization can be utilized to rapidly screen large numbers of other samples for that same mutation. Such a technique can utilize probes which are labeled with gold nanoparticles to yield a visual color result (Elghanian et al., 1997).

In order to detect the alteration of the wild-type HPC2 gene in a tissue, it is helpful to isolate the tissue free from surrounding normal tissues. Means for enriching tissue preparation for tumor cells are known in the art. For example, the tissue may be isolated from paraffin or cryostat sections. Cancer cells may also be separated from normal cells by flow cytometry. These techniques, as well as other techniques for separating tumor cells from normal cells, are

well known in the art. If the tumor tissue is highly contaminated with normal cells, detection of mutations is more difficult.

Detection of point mutations may be accomplished by molecular cloning of the HPC2 allele(s) and sequencing the allele(s) using techniques well known in the art. Alternatively, the gene sequences can be amplified directly from a genomic DNA preparation from the tumor tissue, using known techniques. The DNA sequence of the amplified sequences can then be determined.

There are six well known methods for a more complete, yet still indirect, test for confirming the presence of a susceptibility allele: 1) single-stranded conformation analysis (SSCA) (Orita *et al.*, 1989); 2) denaturing gradient gel electrophoresis (DGGE) (Wartell *et al.*, 1990; Sheffield *et al.*, 1989); 3) RNase protection assays (Finkelstein *et al.*, 1990; Kinszler *et al.*, 1991); 4) allele-specific oligonucleotides (ASOs) (Conner *et al.*, 1983); 5) the use of proteins which recognize nucleotide mismatches, such as the *E. coli* mutS protein (Modrich, 1991); and 6) allele-specific PCR (Ruano and Kidd, 1989). For allele-specific PCR, primers are used which hybridize at their 3' ends to a particular HPC2 mutation. If the particular HPC2 mutation is not present, an amplification product is not observed. Amplification Refractory Mutation System (ARMS) can also be used, as disclosed in European Patent Application Publication No. 0332435 and in Newton *et al.*, 1989. Insertions and deletions of genes can also be detected by cloning, sequencing and amplification. In addition, restriction fragment length polymorphism (RFLP) probes for the gene or surrounding marker genes can be used to score alteration of an allele or an insertion in a polymorphic fragment. Such a method is particularly useful for screening relatives of an affected individual for the presence of the HPC2 mutation found in that individual. Other techniques for detecting insertions and deletions as known in the art can be used.

In the first three methods (SSCA, DGGE and RNase protection assay), a new electrophoretic band appears. SSCA detects a band which migrates differentially because the sequence change causes a difference in single-strand, intramolecular base pairing. RNase protection involves cleavage of the mutant polynucleotide into two or more smaller fragments. DGGE detects differences in migration rates of mutant sequences compared to wild-type sequences, using a denaturing gradient gel. In an allele-specific oligonucleotide assay, an oligonucleotide is designed which detects a specific sequence, and the assay is performed by detecting the presence or absence of a hybridization signal. In the mutS assay, the protein binds

only to sequences that contain a nucleotide mismatch in a heteroduplex between mutant and wild-type sequences.

Mismatches, according to the present invention, are hybridized nucleic acid duplexes in which the two strands are not 100% complementary. Lack of total homology may be due to deletions, insertions, inversions or substitutions. Mismatch detection can be used to detect point mutations in the gene or in its mRNA product. While these techniques are less sensitive than sequencing, they are simpler to perform on a large number of tumor samples. An example of a mismatch cleavage technique is the RNase protection method. In the practice of the present invention, the method involves the use of a labeled riboprobe which is complementary to the human wild-type HPC2 gene coding sequence. The riboprobe and either mRNA or DNA isolated from the tumor tissue are annealed (hybridized) together and subsequently digested with the enzyme RNase A which is able to detect some mismatches in a duplex RNA structure. If a mismatch is detected by RNase A, it cleaves at the site of the mismatch. Thus, when the annealed RNA preparation is separated on an electrophoretic gel matrix, if a mismatch has been detected and cleaved by RNase A, an RNA product will be seen which is smaller than the full length duplex RNA for the riboprobe and the mRNA or DNA. The riboprobe need not be the full length of the HPC2 mRNA or gene but can be a segment of either. If the riboprobe comprises only a segment of the HPC2 mRNA or gene, it will be desirable to use a number of these probes to screen the whole mRNA sequence for mismatches.

In similar fashion, DNA probes can be used to detect mismatches, through enzymatic or chemical cleavage. See, e.g., Cotton *et al.*, 1988; Shenk *et al.*, 1975; Novack *et al.*, 1986. Alternatively, mismatches can be detected by shifts in the electrophoretic mobility of mismatched duplexes relative to matched duplexes. See, e.g., Cariello, 1988. With either riboprobes or DNA probes, the cellular mRNA or DNA which might contain a mutation can be amplified using PCR (see below) before hybridization. Changes in DNA of the HPC2 gene can also be detected using Southern hybridization, especially if the changes are gross rearrangements, such as deletions and insertions.

DNA sequences of the HPC2 gene which have been amplified by use of PCR may also be screened using allele-specific probes. These probes are nucleic acid oligomers, each of which contains a region of the HPC2 gene sequence harboring a known mutation. For example, one oligomer may be about 30 nucleotides in length (although shorter and longer oligomers are also usable as well recognized by those of skill in the art), corresponding to a portion of the HPC2 gene sequence. By use of a battery of such allele-specific probes, PCR amplification

products can be screened to identify the presence of a previously identified mutation in the HPC2 gene. Hybridization of allele-specific probes with amplified HPC2 sequences can be performed, for example, on a nylon filter. Hybridization to a particular probe under high stringency hybridization conditions indicates the presence of the same mutation in the tumor tissue as in the allele-specific probe.

The newly developed technique of nucleic acid analysis via microchip technology is also applicable to the present invention. In this technique, literally thousands of distinct oligonucleotide probes are built up in an array on a silicon chip. Nucleic acid to be analyzed is fluorescently labeled and hybridized to the probes on the chip. It is also possible to study nucleic acid-protein interactions using these nucleic acid microchips. Using this technique one can determine the presence of mutations or even sequence the nucleic acid being analyzed or one can measure expression levels of a gene of interest. The method is one of parallel processing of many, even thousands, of probes at once and can tremendously increase the rate of analysis. Several papers have been published which use this technique. Some of these are Hacia et al., 1996; Shoemaker et al., 1996; Chee et al., 1996; Lockhart et al., 1996; DeRisi et al., 1996; Lipshutz et al., 1995. This method has already been used to screen people for mutations in the breast cancer gene BRCA1 (Hacia et al., 1996). This new technology has been reviewed in a news article in Chemical and Engineering News (Borman, 1996) and been the subject of an editorial (Nature Genetics, 1996). Also see Fodor (1997).

The most definitive test for mutations in a candidate locus is to directly compare genomic HPC2 sequences from cancer patients with those from a control population. Alternatively, one could sequence messenger RNA after amplification, e.g., by PCR, thereby eliminating the necessity of determining the exon structure of the candidate gene.

Mutations from cancer patients falling outside the coding region of HPC2 can be detected by examining the non-coding regions, such as introns and regulatory sequences near or within the HPC2 gene. An early indication that mutations in noncoding regions are important may come from Northern blot experiments that reveal messenger RNA molecules of abnormal size or abundance in cancer patients as compared to control individuals.

Alteration of HPC2 mRNA expression can be detected by any techniques known in the art. These include Northern blot analysis, PCR amplification and RNase protection. Diminished mRNA expression indicates an alteration of the wild-type HPC2 gene. Alteration of wild-type HPC2 genes can also be detected by screening for alteration of wild-type HPC2 protein. For example, monoclonal antibodies immunoreactive with HPC2 can be used to screen a tissue.

Lack of cognate antigen would indicate an HPC2 mutation. Antibodies specific for products of mutant alleles could also be used to detect mutant HPC2 gene product. Such immunological assays can be done in any convenient formats known in the art. These include Western blots, immunohistochemical assays and ELISA assays. Any means for detecting an altered HPC2 protein can be used to detect alteration of wild-type HPC2 genes. Functional assays, such as protein binding determinations, can be used. In addition, assays can be used which detect HPC2 biochemical function. Finding a mutant HPC2 gene product indicates alteration of a wild-type HPC2 gene.

Mutant HPC2 genes or gene products can also be detected in other human body samples, such as serum, stool, urine and sputum. The same techniques discussed above for detection of mutant HPC2 genes or gene products in tissues can be applied to other body samples. Cancer cells are sloughed off from tumors and appear in such body samples. In addition, the HPC2 gene product itself may be secreted into the extracellular space and found in these body samples even in the absence of cancer cells. By screening such body samples, a simple early diagnosis can be achieved for many types of cancers. In addition, the progress of chemotherapy or radiotherapy can be monitored more easily by testing such body samples for mutant HPC2 genes or gene products.

The methods of diagnosis of the present invention are applicable to any tumor in which HPC2 has a role in tumorigenesis. The diagnostic method of the present invention is useful for clinicians, so they can decide upon an appropriate course of treatment.

The primer pairs of the present invention are useful for determination of the nucleotide sequence of a particular HPC2 allele using PCR. The pairs of single-stranded DNA primers can be annealed to sequences within or surrounding the HPC2 gene on chromosome 17 in order to prime amplifying DNA synthesis of the HPC2 gene itself. A complete set of these primers allows synthesis of all of the nucleotides of the HPC2 gene coding sequences, i.e., the exons. The set of primers preferably allows synthesis of both intron and exon sequences. Allele-specific primers can also be used. Such primers anneal only to particular HPC2 mutant alleles, and thus will only amplify a product in the presence of the mutant allele as a template.

In order to facilitate subsequent cloning of amplified sequences, primers may have restriction enzyme site sequences appended to their 5' ends. Thus, all nucleotides of the primers are derived from HPC2 sequences or sequences adjacent to HPC2, except for the few nucleotides necessary to form a restriction enzyme site. Such enzymes and sites are well known in the art. The primers themselves can be synthesized using techniques which are well known in

the art. Generally, the primers can be made using oligonucleotide synthesizing machines which are commercially available. Given the sequence of the HPC2 open reading frame shown in SEQ ID NOs:1 and 3, design of particular primers is well within the skill of the art.

The nucleic acid probes provided by the present invention are useful for a number of purposes. They can be used in Southern hybridization to genomic DNA and in the RNase protection method for detecting point mutations already discussed above. The probes can be used to detect PCR amplification products. They may also be used to detect mismatches with the HPC2 gene or mRNA using other techniques.

It has been discovered that individuals with the wild-type HPC2 gene do not have cancer which results from the HPC2 allele. However, mutations which interfere with the function of the HPC2 protein are involved in the pathogenesis of cancer. Thus, the presence of an altered (or a mutant) HPC2 gene which produces a protein having a loss of function, or altered function, directly correlates to an increased risk of cancer. In order to detect an HPC2 gene mutation, a biological sample is prepared and analyzed for a difference between the sequence of the HPC2 allele being analyzed and the sequence of the wild-type HPC2 allele. Mutant HPC2 alleles can be initially identified by any of the techniques described above. The mutant alleles are then sequenced to identify the specific mutation of the particular mutant allele. Alternatively, mutant HPC2 alleles can be initially identified by identifying mutant (altered) HPC2 proteins, using conventional techniques. The mutant alleles are then sequenced to identify the specific mutation for each allele. The mutations, especially those which lead to an altered function of the HPC2 protein, are then used for the diagnostic and prognostic methods of the present invention.

Definitions

The present invention employs the following definitions:

"Amplification of Polynucleotides" utilizes methods such as the polymerase chain reaction (PCR), ligation amplification (or ligase chain reaction, LCR) and amplification methods based on the use of Q-beta replicase. Also useful are strand displacement amplification (SDA), thermophilic SDA, and nucleic acid sequence based amplification (3SR or NASBA). These methods are well known and widely practiced in the art. See, e.g., U.S. Patents 4,683,195 and 4,683,202 and Innis *et al.*, 1990 (for PCR); and Wu and Wallace, 1989 (for LCR); U.S. Patents 5,270,184 and 5,455,166 and Walker *et al.*, 1992 (for SDA); Spargo et al., 1996 (for thermophilic SDA) and U.S. Patent 5,409,818, Fahy et al., 1991 and Compton, 1991 for 3SR and NASBA. Reagents and hardware for conducting PCR are commercially available. Primers useful to amplify sequences from the HPC2 region or HPC2 paralogs or orthologs are preferably complementary to, and hybridize specifically to sequences in the HPC2 region or paralog or ortholog region or in regions that flank a target region therein. HPC2 sequences or paralog or ortholog sequences generated by amplification may be sequenced directly. Alternatively, but less desirably, the amplified sequence(s) may be cloned prior to sequence analysis. A method for the direct cloning and sequence analysis of enzymatically amplified genomic segments has been described by Scharf, 1986.

"Analyte polynucleotide" and **"analyte strand"** refer to a single- or double-stranded polynucleotide which is suspected of containing a target sequence, and which may be present in a variety of types of samples, including biological samples.

"Antibodies." The present invention also provides polyclonal and/or monoclonal antibodies and fragments thereof, and immunologic binding equivalents thereof, which are capable of specifically binding to the HPC2 polypeptides or to polypeptides encoded by paralogs or orthologs of HPC2 and fragments thereof or to polynucleotide sequences from the HPC2 region, or to polynucleotide sequences which are paralogs or orthologs of HPC2, particularly from the HPC2 locus or a portion thereof. The term **"antibody"** is used both to refer to a homogeneous molecular entity, or a mixture such as a serum product made up of a plurality of different molecular entities. Polypeptides may be prepared synthetically in a peptide synthesizer and coupled to a carrier molecule (e.g., keyhole limpet hemocyanin) and injected over several months into rabbits. Rabbit sera is tested for immunoreactivity to the HPC2 polypeptide or fragment or to polypeptides or fragments encoded by paralogs or orthologs of HPC2. Monoclonal antibodies may be made by injecting mice with the protein polypeptides, fusion

proteins or fragments thereof. Monoclonal antibodies will be screened by ELISA and tested for specific immunoreactivity with HPC2 polypeptide or fragments thereof. See, Harlow and Lane, 1988. These antibodies will be useful in assays as well as pharmaceuticals.

Once a sufficient quantity of desired polypeptide has been obtained, it may be used for various purposes. A typical use is the production of antibodies specific for binding. These antibodies may be either polyclonal or monoclonal, and may be produced by *in vitro* or *in vivo* techniques well known in the art. For production of polyclonal antibodies, an appropriate target immune system, typically mouse or rabbit, is selected. Substantially purified antigen is presented to the immune system in a fashion determined by methods appropriate for the animal and by other parameters well known to immunologists. Typical sites for injection are in footpads, intramuscularly, intraperitoneally, or intradermally. Of course, other species may be substituted for mouse or rabbit. Polyclonal antibodies are then purified using techniques known in the art, adjusted for the desired specificity.

An immunological response is usually assayed with an immunoassay. Normally, such immunoassays involve some purification of a source of antigen, for example, that produced by the same cells and in the same fashion as the antigen. A variety of immunoassay methods are well known in the art. See, e.g., Harlow and Lane, 1988, or Goding, 1986.

Monoclonal antibodies with affinities of 10^{-8} M^{-1} or preferably 10^{-9} to 10^{-10} M^{-1} or stronger will typically be made by standard procedures as described, e.g., in Harlow and Lane, 1988 or Goding, 1986. Briefly, appropriate animals will be selected and the desired immunization protocol followed. After the appropriate period of time, the spleens of such animals are excised and individual spleen cells fused, typically, to immortalized myeloma cells under appropriate selection conditions. Thereafter, the cells are clonally separated and the supernatants of each clone tested for their production of an appropriate antibody specific for the desired region of the antigen.

Other suitable techniques involve *in vitro* exposure of lymphocytes to the antigenic polypeptides, or alternatively, to selection of libraries of antibodies in phage or similar vectors. See Huse *et al.*, 1989. The polypeptides and antibodies of the present invention may be used with or without modification. Frequently, polypeptides and antibodies will be labeled by joining, either covalently or non-covalently, a substance which provides for a detectable signal. A wide variety of labels and conjugation techniques are known and are reported extensively in both the scientific and patent literature. Suitable labels include radionuclides, enzymes, substrates, cofactors, inhibitors, fluorescent agents, chemiluminescent agents, magnetic particles and the

like. Patents teaching the use of such labels include U.S. Patents 3,817,837; 3,850,752; 3,939,350; 3,996,345; 4,277,437; 4,275,149 and 4,366,241. Also, recombinant immunoglobulins may be produced (see U.S. Patent 4,816,567).

"Binding partner" refers to a molecule capable of binding a ligand molecule with high specificity, as for example, an antigen and an antigen-specific antibody or an enzyme and its inhibitor. In general, the specific binding partners must bind with sufficient affinity to immobilize the analyte copy/complementary strand duplex (in the case of polynucleotide hybridization) under the isolation conditions. Specific binding partners are known in the art and include, for example, biotin and avidin or streptavidin, IgG and protein A, the numerous, known receptor-ligand couples, and complementary polynucleotide strands. In the case of complementary polynucleotide binding partners, the partners are normally at least about 15 bases in length, and may be at least 40 bases in length. It is well recognized by those of skill in the art that lengths shorter than 15 (e.g., 8 bases), between 15 and 40, and greater than 40 bases may also be used. The polynucleotides may be composed of DNA, RNA, or synthetic nucleotide analogs. Further binding partners can be identified using, e.g., the two-hybrid yeast screening assay as described herein.

A **"biological sample"** refers to a sample of tissue or fluid suspected of containing an analyte polynucleotide or polypeptide from an individual including, but not limited to, e.g., plasma, serum, spinal fluid, lymph fluid, the external sections of the skin, respiratory, intestinal, and genitourinary tracts, tears, saliva, blood cells, tumors, organs, tissue and samples of *in vitro* cell culture constituents.

As used herein, the terms **"diagnosing"** or **"prognosing,"** as used in the context of neoplasia, are used to indicate 1) the classification of lesions as neoplasia, 2) the determination of the severity of the neoplasia, or 3) the monitoring of the disease progression, prior to, during and after treatment.

"Encode". A polynucleotide is said to "encode" a polypeptide if, in its native state or when manipulated by methods well known to those skilled in the art, it can be transcribed and/or translated to produce the mRNA for and/or the polypeptide or a fragment thereof. The anti-sense strand is the complement of such a nucleic acid, and the encoding sequence can be deduced therefrom.

"Isolated" or **"substantially pure"**. An "isolated" or "substantially pure" nucleic acid (e.g., an RNA, DNA or a mixed polymer) is one which is substantially separated from other cellular components which naturally accompany a native human sequence or protein, e.g.,

ribosomes, polymerases, many other human genome sequences and proteins. The term embraces a nucleic acid sequence or protein which has been removed from its naturally occurring environment, and includes recombinant or cloned DNA isolates and chemically synthesized analogs or analogs biologically synthesized by heterologous systems.

"HPC2 Allele" refers to normal alleles of the HPC2 locus as well as alleles carrying variations that predispose individuals to develop prostate cancer. Such predisposing alleles are also called **"HPC2 susceptibility alleles"**.

"HPC2 Locus", "HPC2 Gene", "HPC2 Nucleic Acids" or **"HPC2 Polynucleotide"** each refer to polynucleotides, all of which are in the HPC2 region, that are likely to be expressed in normal tissue, certain alleles of which predispose an individual to develop prostate cancers. Mutations at the HPC2 locus may be involved in the initiation and/or progression of other types of tumors. The locus is indicated in part by mutations that predispose individuals to develop cancer. These mutations fall within the HPC2 region described *infra*. The HPC2 locus is intended to include coding sequences, intervening sequences and regulatory elements controlling transcription and/or translation. The HPC2 locus is intended to include all allelic variations of the DNA sequence.

The term HPC2 is used interchangeably throughout this disclosure with the terms ELAC2 and HPC2/ELAC2. This holds true regardless of whether the term refers to a nucleic acid, allele, gene, locus, protein or peptide.

These terms, when applied to a nucleic acid, refer to a nucleic acid which encodes an HPC2 polypeptide, fragment, homolog or variant, including, e.g., protein fusions or deletions. The nucleic acids of the present invention will possess a sequence which is either derived from, or substantially similar to a natural HPC2-encoding gene or one having substantial homology with a natural HPC2-encoding gene or a portion thereof.

The HPC2 gene or nucleic acid includes normal alleles of the HPC2 gene, including silent alleles having no effect on the amino acid sequence of the HPC2 polypeptide as well as alleles leading to amino acid sequence variants of the HPC2 polypeptide that do not substantially affect its function. These terms also include alleles having one or more mutations which adversely affect the function of the HPC2 polypeptide. A mutation may be a change in the HPC2 nucleic acid sequence which produces a deleterious change in the amino acid sequence of the HPC2 polypeptide, resulting in partial or complete loss of HPC2 function, or may be a change in the nucleic acid sequence which results in the loss of effective HPC2 expression or the production of aberrant forms of the HPC2 polypeptide.

The HPC2 nucleic acid may be that shown in SEQ ID NOs:1, 3 or 28 or it may be an allele as described above or a variant or derivative differing from that shown by a change which is one or more of addition, insertion, deletion and substitution of one or more nucleotides of the sequence shown. Changes to the nucleotide sequence may result in an amino acid change at the protein level, or not, as determined by the genetic code.

Thus, nucleic acid according to the present invention may include a sequence different from the sequence shown in SEQ ID NOs:1, 3 or 28 yet encode a polypeptide with the same amino acid sequence as shown in SEQ ID NO:1. That is, nucleic acids of the present invention include sequences which are degenerate as a result of the genetic code. On the other hand, the encoded polypeptide may comprise an amino acid sequence which differs by one or more amino acid residues from the amino acid sequence shown in SEQ ID NO:2. Nucleic acid encoding a polypeptide which is an amino acid sequence variant, derivative or allele of the amino acid sequence shown in SEQ ID NO:2 is also provided by the present invention.

The HPC2 gene also refers to (a) any DNA sequence that (i) hybridizes to the complement of the DNA sequences that encode the amino acid sequence set forth in SEQ ID NO:2 under highly stringent conditions (Ausubel et al., 1992) and (ii) encodes a gene product functionally equivalent to HPC2, or (b) any DNA sequence that (i) hybridizes to the complement of the DNA sequences that encode the amino acid sequence set forth in SEQ ID NO:2 under less stringent conditions, such as moderately stringent conditions (Ausubel et al., 1992) and (ii) encodes a gene product functionally equivalent to HPC2. The invention also includes nucleic acid molecules that are the complements of the sequences described herein.

The polynucleotide compositions of this invention include RNA, cDNA, genomic DNA, synthetic forms, and mixed polymers, both sense and antisense strands, and may be chemically or biochemically modified or may contain non-natural or derivatized nucleotide bases, as will be readily appreciated by those skilled in the art. Such modifications include, for example, labels, methylation, substitution of one or more of the naturally occurring nucleotides with an analog, internucleotide modifications such as uncharged linkages (e.g., methyl phosphonates, phosphotriesters, phosphoramidates, carbamates, etc.), charged linkages (e.g., phosphorothioates, phosphorodithioates, etc.), pendent moieties (e.g., polypeptides), intercalators (e.g., acridine, psoralen, etc.), chelators, alkylators, and modified linkages (e.g., alpha anomeric nucleic acids, etc.). Also included are synthetic molecules that mimic polynucleotides in their ability to bind to a designated sequence via hydrogen bonding and other chemical interactions. Such molecules

are known in the art and include, for example, those in which peptide linkages substitute for phosphate linkages in the backbone of the molecule.

The present invention provides recombinant nucleic acids comprising all or part of the HPC2 region or the HPC2 paralog called ELAC1 or the mouse, chimpanzee or gorilla orthologs of HPC2, herein called mouse ELAC2, chimpanzee ELAC2 or gorilla ELAC2. The recombinant construct may be capable of replicating autonomously in a host cell. Alternatively, the recombinant construct may become integrated into the chromosomal DNA of the host cell. Such a recombinant polynucleotide comprises a polynucleotide of genomic, cDNA, semi-synthetic, or synthetic origin which, by virtue of its origin or manipulation, 1) is not associated with all or a portion of a polynucleotide with which it is associated in nature; 2) is linked to a polynucleotide other than that to which it is linked in nature; or 3) does not occur in nature. Where nucleic acid according to the invention includes RNA, reference to the sequence shown should be construed as reference to the RNA equivalent, with U substituted for T.

Therefore, recombinant nucleic acids comprising sequences otherwise not naturally occurring are provided by this invention. Although the wild-type sequence may be employed, it will often be altered, e.g., by deletion, substitution or insertion.

cDNA or genomic libraries of various types may be screened as natural sources of the nucleic acids of the present invention, or such nucleic acids may be provided by amplification of sequences resident in genomic DNA or other natural sources, e.g., by PCR. The choice of cDNA libraries normally corresponds to a tissue source which is abundant in mRNA for the desired proteins. Phage libraries are normally preferred, but other types of libraries may be used. Clones of a library are spread onto plates, transferred to a substrate for screening, denatured and probed for the presence of desired sequences.

The DNA sequences used in this invention will usually comprise at least about five codons (15 nucleotides), more usually at least about 7-15 codons, and most preferably, at least about 35 codons. One or more introns may also be present. This number of nucleotides is usually about the minimal length required for a successful probe that would hybridize specifically with an HPC2-encoding sequence. In this context, oligomers of as low as 8 nucleotides, more generally 8-17 nucleotides, can be used for probes, especially in connection with chip technology.

Techniques for nucleic acid manipulation are described generally, for example, in Sambrook *et al.*, 1989 or Ausubel *et al.*, 1992. Reagents useful in applying such techniques, such as restriction enzymes and the like, are widely known in the art and commercially available

from such vendors as New England BioLabs, Boehringer Mannheim, Amersham, Promega Biotec, U. S. Biochemicals, New England Nuclear, and a number of other sources. The recombinant nucleic acid sequences used to produce fusion proteins of the present invention may be derived from natural or synthetic sequences. Many natural gene sequences are obtainable from various cDNA or from genomic libraries using appropriate probes. See, GenBank, National Institutes of Health.

"HPC2 Region" refers to a portion of human chromosome 17 bounded by the markers D17S947 and D17S799. This region contains the HPC2 locus, including the HPC2 gene.

As used herein, the terms **"HPC2 locus", "HPC2 allele"** and **"HPC2 region"** all refer to the double-stranded DNA comprising the locus, allele, or region, as well as either of the single-stranded DNAs comprising the locus, allele or region.

As used herein, a **"portion"** of the *HPC2* locus or region or allele is defined as having a minimal size of at least about eight nucleotides, or preferably about 15 nucleotides, or more preferably at least about 25 nucleotides, and may have a minimal size of at least about 40 nucleotides. This definition includes all sizes in the range of 8-40 nucleotides as well as greater than 40 nucleotides. Thus, this definition includes nucleic acids of 8, 12, 15, 20, 25, 40, 60, 80, 100, 200, 300, 400, 500 nucleotides, or nucleic acids having any number of nucleotides within these ranges of values (e.g., 9, 10, 11, 16, 23, 30, 38, 50, 72, 121, etc., nucleotides), or nucleic acids having more than 500 nucleotides. The present invention includes all novel nucleic acids having at least 8 nucleotides derived from any of SEQ ID NOs:1 or 3-28, its complement or functionally equivalent nucleic acid sequences. The present invention does not include nucleic acids which exist in the prior art. That is, the present invention includes all nucleic acids having at least 8 nucleotides derived from any of SEQ ID NOs:1 or 3-28 with the proviso that it does not include nucleic acids existing in the prior art.

"HPC2 protein" or **"HPC2 polypeptide"** refers to a protein or polypeptide encoded by the HPC2 locus, variants or fragments thereof. The term "polypeptide" refers to a polymer of amino acids and its equivalent and does not refer to a specific length of the product; thus, peptides, oligopeptides and proteins are included within the definition of a polypeptide. This term also does not refer to, or exclude modifications of the polypeptide, for example, glycosylations, acetylations, phosphorylations, and the like. Included within the definition are, for example, polypeptides containing one or more analogs of an amino acid (including, for example, unnatural amino acids, etc.), polypeptides with substituted linkages as well as other modifications known in the art, both naturally and non-naturally occurring. Ordinarily, such

polypeptides will be at least about 50% homologous to the native HPC2 sequence, preferably in excess of about 90%, and more preferably at least about 95% homologous. Also included are proteins encoded by DNA which hybridize under high or low stringency conditions, to HPC2-encoding nucleic acids and closely related polypeptides or proteins retrieved by antisera to the HPC2 protein(s).

An HPC2 polypeptide may be that derived from any of the exons described herein which may be in isolated and/or purified form, free or substantially free of material with which it is naturally associated. The polypeptide may, if produced by expression in a prokaryotic cell or produced synthetically, lack native post-translational processing, such as glycosylation. Alternatively, the present invention is also directed to polypeptides which are sequence variants, alleles or derivatives of an HPC2 polypeptide. Such polypeptides may have an amino acid sequence which differs from that derived from any of the exons described herein by one or more of addition, substitution, deletion or insertion of one or more amino acids. Preferred such polypeptides have HPC2 function.

Substitutional variants typically contain the exchange of one amino acid for another at one or more sites within the protein, and may be designed to modulate one or more properties of the polypeptide, such as stability against proteolytic cleavage, without the loss of other functions or properties. Amino acid substitutions may be made on the basis of similarity in polarity, charge, solubility, hydrophobicity, hydrophilicity, and/or the amphipathic nature of the residues involved. Preferred substitutions are ones which are conservative, that is, one amino acid is replaced with one of similar shape and charge. Conservative substitutions are well known in the art and typically include substitutions within the following groups: glycine, alanine; valine, isoleucine, leucine; aspartic acid, glutamic acid; asparagine, glutamine; serine, threonine; lysine, arginine; and tyrosine, phenylalanine.

Certain amino acids may be substituted for other amino acids in a protein structure without appreciable loss of interactive binding capacity with structures such as, for example, antigen-binding regions of antibodies or binding sites on substrate molecules or binding sites on proteins interacting with an HPC2 polypeptide. Since it is the interactive capacity and nature of a protein which defines that protein's biological functional activity, certain amino acid substitutions can be made in a protein sequence, and its underlying DNA coding sequence, and nevertheless obtain a protein with like properties. In making such changes, the hydropathic index of amino acids may be considered. The importance of the hydrophobic amino acid index in conferring interactive biological function on a protein is generally understood in the art (Kyte

and Doolittle, 1982). Alternatively, the substitution of like amino acids can be made effectively on the basis of hydrophilicity. The importance of hydrophilicity in conferring interactive biological function of a protein is generally understood in the art (U.S. Patent 4,554,101). The use of the hydrophobic index or hydrophilicity in designing polypeptides is further discussed in U.S. Patent 5,691,198.

The length of polypeptide sequences compared for homology will generally be at least about 16 amino acids, usually at least about 20 residues, more usually at least about 24 residues, typically at least about 28 residues, and preferably more than about 35 residues.

"Operably linked" refers to a juxtaposition wherein the components so described are in a relationship permitting them to function in their intended manner. For instance, a promoter is operably linked to a coding sequence if the promoter affects its transcription or expression.

The term **peptide mimetic** or **mimetic** is intended to refer to a substance which has the essential biological activity of an HPC2, ELAC1 or ELAC2 polypeptide. A peptide mimetic may be a peptide-containing molecule that mimics elements of protein secondary structure (Johnson et al., 1993). The underlying rationale behind the use of peptide mimetics is that the peptide backbone of proteins exists chiefly to orient amino acid side chains in such a way as to facilitate molecular interactions, such as those of antibody and antigen, enzyme and substrate or scaffolding proteins. A peptide mimetic is designed to permit molecular interactions similar to the natural molecule. A mimetic may not be a peptide at all, but it will retain the essential biological activity of a natural HPC2, ELAC1 or ELAC2 polypeptide.

"Probes". Polynucleotide polymorphisms associated with HPC2 alleles which predispose to certain cancers or are associated with most cancers are detected by hybridization with a polynucleotide probe which forms a stable hybrid with that of the target sequence, under highly stringent to moderately stringent hybridization and wash conditions. If it is expected that the probes will be perfectly complementary to the target sequence, high stringency conditions will be used. Hybridization stringency may be lessened if some mismatching is expected, for example, if variants are expected with the result that the probe will not be completely complementary. Conditions are chosen which rule out nonspecific/adventitious bindings, that is, which minimize noise. (It should be noted that throughout this disclosure, if it is simply stated that "stringent" conditions are used that is meant to be read as "high stringency" conditions are used.) Since such indications identify neutral DNA polymorphisms as well as mutations, these indications need further analysis to demonstrate detection of an HPC2 susceptibility allele. An example of high stringency conditions is to hybridize to filter bound DNA in 0.5 M $NaHPO_4$,

7% sodium dodecyl sulfate (SDS), 1 mM EDTA at 65°C and to wash in 0.1xSSC/0.1% SDS at 68°C (Ausubel et al., 1992). Less stringent conditions, such as moderately stringent conditions, are defined as above but with the wash step being in 0.2xSSC/0.1% SDS at 42°C.

Probes for HPC2 alleles may be derived from the sequences of the HPC2 region, its cDNA, functionally equivalent sequences, or the complements thereof. The probes may be of any suitable length, which span all or a portion of the HPC2 region, and which allow specific hybridization to the HPC2 region. If the target sequence contains a sequence identical to that of the probe, the probes may be short, e.g., in the range of about 8-30 base pairs, since the hybrid will be relatively stable under even highly stringent conditions. If some degree of mismatch is expected with the probe, i.e., if it is suspected that the probe will hybridize to a variant region, a longer probe may be employed which hybridizes to the target sequence with the requisite specificity.

The probes will include an isolated polynucleotide attached to a label or reporter molecule and may be used to isolate other polynucleotide sequences, having sequence similarity by standard methods. For techniques for preparing and labeling probes see, e.g., Sambrook *et al.*, 1989 or Ausubel *et al.*, 1992. Other similar polynucleotides may be selected by using homologous polynucleotides. Alternatively, polynucleotides encoding these or similar polypeptides may be synthesized or selected by use of the redundancy in the genetic code. Various codon substitutions may be introduced, e.g., by silent changes (thereby producing various restriction sites) or to optimize expression for a particular system. Mutations may be introduced to modify the properties of the polypeptide, perhaps to change ligand-binding affinities, interchain affinities, or the polypeptide degradation or turnover rate.

Probes comprising synthetic oligonucleotides or other polynucleotides of the present invention may be derived from naturally occurring or recombinant single- or double-stranded polynucleotides, or be chemically synthesized. Probes may also be labeled by nick translation, Klenow fill-in reaction, or other methods known in the art.

Portions of the polynucleotide sequence having at least about eight nucleotides, usually at least about 15 nucleotides, and fewer than about 9 kb, usually fewer than about 1.0 kb, from a polynucleotide sequence encoding HPC2 are preferred as probes. This definition therefore includes probes of sizes 8 nucleotides through 9000 nucleotides. Thus, this definition includes probes of 8, 12, 15, 20, 25, 40, 60, 80, 100, 200, 300, 400 or 500 nucleotides or probes having any number of nucleotides within these ranges of values (e.g., 9, 10, 11, 16, 23, 30, 38, 50, 72, 121, etc., nucleotides), or probes having more than 500 nucleotides. The probes may also be

used to determine whether mRNA encoding HPC2 is present in a cell or tissue. The present invention includes all novel probes having at least 8 nucleotides derived from any of SEQ ID NOs:1 or 3-28 its complement or functionally equivalent nucleic acid sequences. The present invention does not include probes which exist in the prior art. That is, the present invention includes all probes having at least 8 nucleotides derived from any of SEQ ID NOs:1 or 3-28 with the proviso that they do not include probes existing in the prior art.

Similar considerations and nucleotide lengths are also applicable to primers which may be used for the amplification of all or part of the *HPC2* gene. Thus, a definition for primers includes primers of 8, 12, 15, 20, 25, 40, 60, 80, 100, 200, 300, 400, 500 nucleotides, or primers having any number of nucleotides within these ranges of values (e.g., 9, 10, 11, 16, 23, 30, 38, 50, 72, 121, etc. nucleotides), or primers having more than 500 nucleotides, or any number of nucleotides between 500 and 9000. The primers may also be used to determine whether mRNA encoding *HPC2* is present in a cell or tissue. The present invention includes all novel primers having at least 8 nucleotides derived from the *HPC2* locus for amplifying the *HPC2* gene, its complement or functionally equivalent nucleic acid sequences. The present invention does not include primers which exist in the prior art. That is, the present invention includes all primers having at least 8 nucleotides with the proviso that it does not include primers existing in the prior art.

"Protein modifications or fragments" are provided by the present invention for HPC2, ELAC1 and ELAC2 polypeptides or fragments thereof which are substantially homologous to primary structural sequence but which include, e.g., *in vivo* or *in vitro* chemical and biochemical modifications or which incorporate unusual amino acids. Such modifications include, for example, acetylation, carboxylation, phosphorylation, glycosylation, ubiquitination, labeling, e.g., with radionuclides, and various enzymatic modifications, as will be readily appreciated by those well skilled in the art. A variety of methods for labeling polypeptides and of substituents or labels useful for such purposes are well known in the art, and include radioactive isotopes such as ^{32}P, ligands which bind to labeled antiligands (e.g., antibodies), fluorophores, chemiluminescent agents, enzymes, and antiligands which can serve as specific binding pair members for a labeled ligand. The choice of label depends on the sensitivity required, ease of conjugation with the primer, stability requirements, and available instrumentation. Methods of labeling polypeptides are well known in the art. See Sambrook *et al.*, 1989 or Ausubel *et al.*, 1992.

Besides substantially full-length polypeptides, the present invention provides for biologically active fragments of the polypeptides. Significant biological activities include ligand-binding, immunological activity and other biological activities characteristic of HPC2, ELAC1 or ELAC2 polypeptides. Immunological activities include both immunogenic function in a target immune system, as well as sharing of immunological epitopes for binding, serving as either a competitor or substitute antigen for an epitope of the HPC2, ELAC1 or ELAC2 protein. As used herein, "epitope" refers to an antigenic determinant of a polypeptide. An epitope could comprise three amino acids in a spatial conformation which is unique to the epitope. Generally, an epitope consists of at least five such amino acids, and more usually consists of at least 8-10 such amino acids. Methods of determining the spatial conformation of such amino acids are known in the art.

For immunological purposes, tandem-repeat polypeptide segments may be used as immunogens, thereby producing highly antigenic proteins. Alternatively, such polypeptides will serve as highly efficient competitors for specific binding. Production of antibodies specific for HPC2, ELAC1 or ELAC2 polypeptides or fragments thereof is described below.

The present invention also provides for fusion polypeptides, comprising HPC2, ELAC1 or ELAC2 polypeptides and fragments. Homologous polypeptides may be fusions between two or more HPC2, ELAC1 or ELAC2 polypeptide sequences or between the sequences of HPC2, ELAC1 or ELAC2 and a related protein. Likewise, heterologous fusions may be constructed which would exhibit a combination of properties or activities of the derivative proteins. For example, ligand-binding or other domains may be "swapped" between different new fusion polypeptides or fragments. Such homologous or heterologous fusion polypeptides may display, for example, altered strength or specificity of binding. Fusion partners include immunoglobulins, bacterial β-galactosidase, trpE, protein A, β-lactamase, alpha amylase, alcohol dehydrogenase and yeast alpha mating factor. See Godowski *et al.*, 1988.

Fusion proteins will typically be made by either recombinant nucleic acid methods, as described below, or may be chemically synthesized. Techniques for the synthesis of polypeptides are described, for example, in Merrifield, 1963.

"Protein purification" refers to various methods for the isolation of the HPC2, ELAC1 or ELAC2 polypeptides from other biological material, such as from cells transformed with recombinant nucleic acids encoding HPC2, ELAC1 or ELAC2 and are well known in the art. For example, such polypeptides may be purified by immunoaffinity chromatography employing,

e.g., the antibodies provided by the present invention. Various methods of protein purification are well known in the art, and include those described in Deutscher, 1990 and Scopes, 1982.

The terms **"isolated"**, **"substantially pure"**, and **"substantially homogeneous"** are used interchangeably to describe a protein or polypeptide which has been separated from components which accompany it in its natural state. A monomeric protein is substantially pure when at least about 60 to 75% of a sample exhibits a single polypeptide sequence. A substantially pure protein will typically comprise about 60 to 90% W/W of a protein sample, more usually about 95%, and preferably will be over about 99% pure. Protein purity or homogeneity may be indicated by a number of means well known in the art, such as polyacrylamide gel electrophoresis of a protein sample, followed by visualizing a single polypeptide band upon staining the gel. For certain purposes, higher resolution may be provided by using HPLC or other means well known in the art which are utilized for purification.

An HPC2, ELAC1 or ELAC2 protein is substantially free of naturally associated components when it is separated from the native contaminants which accompany it in its natural state. Thus, a polypeptide which is chemically synthesized or synthesized in a cellular system different from the cell from which it naturally originates will be substantially free from its naturally associated components. A protein may also be rendered substantially free of naturally associated components by isolation, using protein purification techniques well known in the art.

A polypeptide produced as an expression product of an isolated and manipulated genetic sequence is an "isolated polypeptide," as used herein, even if expressed in a homologous cell type. Synthetically made forms or molecules expressed by heterologous cells are inherently isolated molecules.

"Recombinant nucleic acid" is a nucleic acid which is not naturally occurring, or which is made by the artificial combination of two otherwise separated segments of sequence. This artificial combination is often accomplished by either chemical synthesis means, or by the artificial manipulation of isolated segments of nucleic acids, e.g., by genetic engineering techniques. Such is usually done to replace a codon with a redundant codon encoding the same or a conservative amino acid, while typically introducing or removing a sequence recognition site. Alternatively, it is performed to join together nucleic acid segments of desired functions to generate a desired combination of functions.

"Regulatory sequences" refers to those sequences normally within 100 kb of the coding region of a locus, but they may also be more distant from the coding region, which affect the

expression of the gene (including transcription of the gene, and translation, splicing, stability or the like of the messenger RNA).

"Substantial homology or similarity". A nucleic acid or fragment thereof is "substantially homologous" ("or substantially similar") to another if, when optimally aligned (with appropriate nucleotide insertions or deletions) with the other nucleic acid (or its complementary strand), there is nucleotide sequence identity in at least about 60% of the nucleotide bases, usually at least about 70%, more usually at least about 80%, preferably at least about 90%, and more preferably at least about 95-98% of the nucleotide bases.

Identity means the degree of sequence relatedness between two polypeptide or two polynucleotides sequences as determined by the identity of the match between two strings of such sequences. Identity can be readily calculated. While there exist a number of methods to measure identity between two polynucleotide or polypeptide sequences, the term "identity" is well known to skilled artisans (Computational Molecular Biology, Lesk, A. M., ed., Oxford University Press, New York, 1988; Biocomputing: Informatics and Genome Projects, Smith, D. W., ed., Academic Press, New York, 1993; Computer Analysis of Sequence Data, Part I, Griffin, A. M., and Griffin, H. G., eds., Humana Press, New Jersey, 1994; Sequence Analysis in Molecular Biology, von Heinje, G., Academic Press, 1987; and Sequence Analysis Primer, Gribskov, M. and Devereux, J., eds., M Stockton Press, New York, 1991). Methods commonly employed to determine identity between two sequences include, but are not limited to those disclosed in Guide to Huge Computers, Martin J. Bishop, ed., Academic Press, San Diego, 1994, and Carillo, H., and Lipman, D. (1988). Preferred methods to determine identity are designed to give the largest match between the two sequences tested. Such methods are codified in computer programs. Preferred computer program methods to determine identity between two sequences include, but are not limited to, GCG program package (Devereux et al. (1984), BLASTP, BLASTN, FASTA (Altschul et al. (1990); Altschul et al. (1997)).

Alternatively, substantial homology or similarity exists when a nucleic acid or fragment thereof will hybridize to another nucleic acid (or a complementary strand thereof) under selective hybridization conditions, to a strand, or to its complement. Selectivity of hybridization exists when hybridization which is substantially more selective than total lack of specificity occurs. Typically, selective hybridization will occur when there is at least about 55% homology over a stretch of at least about 14 nucleotides, preferably at least about 65%, more preferably at least about 75%, and most preferably at least about 90%. See, Kanehisa, 1984. The length of homology comparison, as described, may be over longer stretches, and in certain embodiments

will often be over a stretch of at least about nine nucleotides, usually at least about 20 nucleotides, more usually at least about 24 nucleotides, typically at least about 28 nucleotides, more typically at least about 32 nucleotides, and preferably at least about 36 or more nucleotides.

Nucleic acid hybridization will be affected by such conditions as salt concentration, temperature, or organic solvents, in addition to the base composition, length of the complementary strands, and the number of nucleotide base mismatches between the hybridizing nucleic acids, as will be readily appreciated by those skilled in the art. Stringent temperature conditions will generally include temperatures in excess of $30^{o}C$, typically in excess of $37^{o}C$, and preferably in excess of $45^{o}C$. Stringent salt conditions will ordinarily be less than 1000 mM, typically less than 500 mM, and preferably less than 200 mM. However, the combination of parameters is much more important than the measure of any single parameter. See, e.g., Wetmur and Davidson, 1968.

Probe sequences may also hybridize specifically to duplex DNA under certain conditions to form triplex or other higher order DNA complexes. The preparation of such probes and suitable hybridization conditions are well known in the art.

The terms **"substantial homology"** or **"substantial identity"**, when referring to polypeptides, indicate that the polypeptide or protein in question exhibits at least about 30% identity with an entire naturally-occurring protein or a portion thereof, usually at least about 70% identity, more usually at least about 80% identity, preferably at least about 90% identity, and more preferably at least about 95% identity.

Homology, for polypeptides, is typically measured using sequence analysis software. See, e.g., the Sequence Analysis Software Package of the Genetics Computer Group, University of Wisconsin Biotechnology Center, 910 University Avenue, Madison, Wisconsin 53705, as well as the software described above with reference to nucleic acid homology. Protein analysis software matches similar sequences using measures of homology assigned to various substitutions, deletions and other modifications. Conservative substitutions typically include substitutions within the following groups: glycine, alanine; valine, isoleucine, leucine; aspartic acid, glutamic acid; asparagine, glutamine; serine, threonine; lysine, arginine; and phenylalanine, tyrosine.

"Substantially similar function" refers to the function of a modified nucleic acid or a modified protein, with reference to the wild-type HPC2, ELAC1 or ELAC2 nucleic acid or wild-type HPC2, ELAC1 or ELAC2 polypeptide. The modified polypeptide will be substantially

homologous to the wild-type HPC2, ELAC1 or ELAC2 polypeptide and will have substantially the same function. The modified polypeptide may have an altered amino acid sequence and/or may contain modified amino acids. In addition to the similarity of function, the modified polypeptide may have other useful properties, such as a longer half-life. The similarity of function (activity) of the modified polypeptide may be substantially the same as the activity of the wild-type HPC2, ELAC1 or ELAC2 polypeptide. Alternatively, the similarity of function (activity) of the modified polypeptide may be higher than the activity of the wild-type HPC2, ELAC1 or ELAC2 polypeptide. The modified polypeptide is synthesized using conventional techniques, or is encoded by a modified nucleic acid and produced using conventional techniques. The modified nucleic acid is prepared by conventional techniques. A nucleic acid with a function substantially similar to the wild-type HPC2, ELAC1 or ELAC2 gene function produces the modified protein described above.

A polypeptide "**fragment**," "**portion**" or "**segment**" is a stretch of amino acid residues of at least about five to seven contiguous amino acids, often at least about seven to nine contiguous amino acids, typically at least about nine to 13 contiguous amino acids and, most preferably, at least about 20 to 30 or more contiguous amino acids.

The polypeptides of the present invention, if soluble, may be coupled to a solid-phase support, e.g., nitrocellulose, nylon, column packing materials (e.g., Sepharose beads), magnetic beads, glass wool, plastic, metal, polymer gels, cells, or other substrates. Such supports may take the form, for example, of beads, wells, dipsticks, or membranes.

"**Target region**" refers to a region of the nucleic acid which is amplified and/or detected. The term "**target sequence**" refers to a sequence with which a probe or primer will form a stable hybrid under desired conditions.

The practice of the present invention employs, unless otherwise indicated, conventional techniques of chemistry, molecular biology, microbiology, recombinant DNA, genetics, and immunology. See, e.g., Maniatis *et al.*, 1982; Sambrook *et al.*, 1989; Ausubel *et al.*, 1992; Glover, 1985; Anand, 1992; Guthrie and Fink, 1991. A general discussion of techniques and materials for human gene mapping, including mapping of human chromosome 1, is provided, e.g., in White and Lalouel, 1988.

Preparation of recombinant or chemically
synthesized nucleic acids; vectors, transformation, host cells

Large amounts of the polynucleotides of the present invention may be produced by replication in a suitable host cell. Natural or synthetic polynucleotide fragments coding for a desired fragment will be incorporated into recombinant polynucleotide constructs, usually DNA constructs, capable of introduction into and replication in a prokaryotic or eukaryotic cell. Usually the polynucleotide constructs will be suitable for replication in a unicellular host, such as yeast or bacteria, but may also be intended for introduction to (with and without integration within the genome) cultured mammalian or plant or other eukaryotic cell lines. The purification of nucleic acids produced by the methods of the present invention is described, e.g., in Sambrook *et al.*, 1989 or Ausubel *et al.*, 1992.

The polynucleotides of the present invention may also be produced by chemical synthesis, e.g., by the phosphoramidite method described by Beaucage and Caruthers, 1981 or the triester method according to Matteucci and Caruthers, 1981, and may be performed on commercial, automated oligonucleotide synthesizers. A double-stranded fragment may be obtained from the single-stranded product of chemical synthesis either by synthesizing the complementary strand and annealing the strands together under appropriate conditions or by adding the complementary strand using DNA polymerase with an appropriate primer sequence.

Polynucleotide constructs prepared for introduction into a prokaryotic or eukaryotic host may comprise a replication system recognized by the host, including the intended polynucleotide fragment encoding the desired polypeptide, and will preferably also include transcription and translational initiation regulatory sequences operably linked to the polypeptide encoding segment. Expression vectors may include, for example, an origin of replication or autonomously replicating sequence (ARS) and expression control sequences, a promoter, an enhancer and necessary processing information sites, such as ribosome-binding sites, RNA splice sites, polyadenylation sites, transcriptional terminator sequences, and mRNA stabilizing sequences. Secretion signals may also be included where appropriate, whether from a native HPC2 protein or from other receptors or from secreted polypeptides of the same or related species, which allow the protein to cross and/or lodge in cell membranes, and thus attain its functional topology, or be secreted from the cell. Such vectors may be prepared by means of standard recombinant techniques well known in the art and discussed, for example, in Sambrook *et al.*, 1989 or Ausubel *et al.* 1992.

An appropriate promoter and other necessary vector sequences will be selected so as to be functional in the host, and may include, when appropriate, those naturally associated with HPC2, ELAC1 or ELAC2 genes. Examples of workable combinations of cell lines and expression vectors are described in Sambrook *et al.*, 1989 or Ausubel *et al.*, 1992; see also, e.g., Metzger *et al.*, 1988. Many useful vectors are known in the art and may be obtained from such vendors as Stratagene, New England BioLabs, Promega Biotech, and others. Promoters such as the trp, lac and phage promoters, tRNA promoters and glycolytic enzyme promoters may be used in prokaryotic hosts. Useful yeast promoters include promoter regions for metallothionein, 3-phosphoglycerate kinase or other glycolytic enzymes such as enolase or glyceraldehyde-3-phosphate dehydrogenase, enzymes responsible for maltose and galactose utilization, and others. Vectors and promoters suitable for use in yeast expression are further described in Hitzeman *et al.*, EP 73,675A. Appropriate non-native mammalian promoters might include the early and late promoters from SV40 (Fiers *et al.*, 1978) or promoters derived from murine Moloney leukemia virus, mouse tumor virus, avian sarcoma viruses, adenovirus II, bovine papilloma virus or polyoma. Insect promoters may be derived from baculovirus. In addition, the construct may be joined to an amplifiable gene (e.g., DHFR) so that multiple copies of the gene may be made. For appropriate enhancer and other expression control sequences, see also *Enhancers and Eukaryotic Gene Expression*, Cold Spring Harbor Press, Cold Spring Harbor, New York (1983). See also, e.g., U.S. Patent Nos. 5,691,198; 5,735,500; 5,747,469 and 5,436,146.

While such expression vectors may replicate autonomously, they may also replicate by being inserted into the genome of the host cell, by methods well known in the art.

Expression and cloning vectors will likely contain a selectable marker, a gene encoding a protein necessary for survival or growth of a host cell transformed with the vector. The presence of this gene ensures growth of only those host cells which express the inserts. Typical selection genes encode proteins that a) confer resistance to antibiotics or other toxic substances, e.g. ampicillin, neomycin, methotrexate, etc.; b) complement auxotrophic deficiencies, or c) supply critical nutrients not available from complex media, e.g., the gene encoding D-alanine racemase for *Bacilli*. The choice of the proper selectable marker will depend on the host cell, and appropriate markers for different hosts are well known in the art.

The vectors containing the nucleic acids of interest can be transcribed *in vitro*, and the resulting RNA introduced into the host cell by well-known methods, e.g., by injection (see, Kubo *et al.*, 1988), or the vectors can be introduced directly into host cells by methods well known in the art, which vary depending on the type of cellular host, including electroporation;

transfection employing calcium chloride, rubidium chloride, calcium phosphate, DEAE-dextran, or other substances; microprojectile bombardment; lipofection; infection (where the vector is an infectious agent, such as a retroviral genome); and other methods. See generally, Sambrook *et al.*, 1989 and Ausubel *et al.*, 1992. The introduction of the polynucleotides into the host cell by any method known in the art, including, *inter alia*, those described above, will be referred to herein as "transformation." The cells into which have been introduced nucleic acids described above are meant to also include the progeny of such cells.

Large quantities of the nucleic acids and polypeptides of the present invention may be prepared by expressing the HPC2, ELAC1 or ELAC2 nucleic acids or portions thereof in vectors or other expression vehicles in compatible prokaryotic or eukaryotic host cells. The most commonly used prokaryotic hosts are strains of *Escherichia coli*, although other prokaryotes, such as *Bacillus subtilis* or *Pseudomonas* may also be used.

Mammalian or other eukaryotic host cells, such as those of yeast, filamentous fungi, plant, insect, or amphibian or avian species, may also be useful for production of the proteins of the present invention. Propagation of mammalian cells in culture is *per se* well known. See, Jakoby and Pastan, 1979. Examples of commonly used mammalian host cell lines are VERO and HeLa cells, Chinese hamster ovary (CHO) cells, and WI38, BHK, and COS cell lines, although it will be appreciated by the skilled practitioner that other cell lines may be appropriate, e.g., to provide higher expression, desirable glycosylation patterns, or other features. An example of a commonly used insect cell line is SF9.

Clones are selected by using markers depending on the mode of the vector construction. The marker may be on the same or a different DNA molecule, preferably the same DNA molecule. In prokaryotic hosts, the transformant may be selected, e.g., by resistance to ampicillin, tetracycline or other antibiotics. Production of a particular product based on temperature sensitivity may also serve as an appropriate marker.

Prokaryotic or eukaryotic cells transformed with the polynucleotides of the present invention will be useful not only for the production of the nucleic acids and polypeptides of the present invention, but also, for example, in studying the characteristics of HPC2, ELAC1 or ELAC2 polypeptides.

The HPC2, ELAC1 or ELAC2 gene products can also be expressed in transgenic animals. Animals of any species, including, but not limited to, mice, rats, rabbits, guinea pigs, pigs, micro-pigs, goats and non-human primates, e.g., baboons, monkeys and chimpanzees, may be used to generate HPC2, ELAC1 or ELAC2 transgenic animals.

Any technique known in the art may be used to introduce the HPC2, ELAC1 or ELAC2 gene transgene into animals to produce the founder lines of transgenic animals. Such techniques include, but are not limited to, pronuclear microinjection (U.S. Patent No. 4,873,191); retrovirus mediated gene transfer into germ lines (Van der Putten et al., 1985); gene targeting in embryonic stem cells (Thompson et al., 1989); electroporation of embryos (Lo, 1983); and sperm-mediated gene transfer (Lavitrano et al., 1989); etc. For a review of such techniques, see Gordon (1989), which is incorporated by reference herein in its entirety.

The present invention provides for transgenic animals that carry the HPC2, ELAC1 or ELAC2 transgene in all their cells, as well as animals which carry the transgene in some, but not all of their cells, i.e., mosaic animals. The transgene may be integrated as a single transgene or in concatamers, e.g., head-to-head tandems or head-to-tail tandems. The transgene may also be selectively introduced into and activated in a particular cell type by following, for example, the teaching of Lasko et al. (1992). The regulatory sequences required for such a cell-type specific activation will depend upon the particular cell type of interest, and will be apparent to those of skill in the art. When it is desired that the HPC2, ELAC1 or ELAC2 gene transgene be integrated into the chromosomal site of the endogenous HPC2, ELAC1 or ELAC2 gene, gene targeting is preferred. Briefly, when such a technique is to be utilized, vectors containing some nucleotide sequences homologous to the endogenous HPC2, ELAC1 or ELAC2 gene are designed for the purpose of integrating, via homologous recombination with chromosomal sequences, into and disrupting the function of the nucleotide sequence of the endogenous HPC2, ELAC1 or ELAC2 gene. The transgene may also be selectively introduced into a particular cell type, thus inactivating the endogenous HPC2, ELAC1 or ELAC2 gene in only that cell type, by following, for example, the teaching of Gu et al. (1994). The regulatory sequences required for such a cell-type specific inactivation will depend upon the particular cell type of interest, and will be apparent to those of skill in the art.

Once transgenic animals have been generated, the expression of the recombinant HPC2, ELAC1 or ELAC2 gene may be assayed utilizing standard techniques. Initial screening may be accomplished by Southern blot analysis or PCR techniques to analyze animal tissues to assay whether integration of the transgene has taken place. The level of mRNA expression of the transgene in the tissues of the transgenic animals may also be assessed using techniques which include, but are not limited to, Northern blot analysis of tissue samples obtained from the animal, *in situ* hybridization analysis, and RT-PCR. Samples of HPC2, ELAC1 or ELAC2

gene-expressing tissue, may also be evaluated immunocytochemically using antibodies specific for the HPC2, ELAC1 or ELAC2 transgene product.

Antisense polynucleotide sequences are useful in preventing or diminishing the expression of the HPC2, ELAC1 or ELAC2 locus, as will be appreciated by those skilled in the art. For example, polynucleotide vectors containing all or a portion of the HPC2 locus or other sequences from the HPC2 region (particularly those flanking the HPC2 locus) may be placed under the control of a promoter in an antisense orientation and introduced into a cell. Expression of such an antisense construct within a cell will interfere with HPC2 transcription and/or translation and/or replication.

The probes and primers based on the HPC2 gene sequences disclosed herein are used to identify homologous HPC2 gene sequences and proteins in other species. These HPC2 gene sequences and proteins are used in the diagnostic/prognostic, therapeutic and drug screening methods described herein for the species from which they have been isolated.

Methods of Use: Nucleic Acid Diagnosis and Diagnostic Kits

In order to detect the presence of an HPC2 allele predisposing an individual to cancer, a biological sample such as blood is prepared and analyzed for the presence or absence of susceptibility alleles of HPC2. In order to detect the presence of neoplasia, the progression toward malignancy of a precursor lesion, or as a prognostic indicator, a biological sample of the lesion is prepared and analyzed for the presence or absence of mutant alleles of HPC2. Results of these tests and interpretive information are returned to the health care provider for communication to the tested individual. Such diagnoses may be performed by diagnostic laboratories, or, alternatively, diagnostic kits are manufactured and sold to health care providers or to private individuals for self-diagnosis.

Initially, the screening method involves amplification of the relevant HPC2 sequences. In another preferred embodiment of the invention, the screening method involves a non-PCR based strategy. Such screening methods include two-step label amplification methodologies that are well known in the art. Both PCR and non-PCR based screening strategies can detect target sequences with a high level of sensitivity.

The most popular method used today is target amplification. Here, the target nucleic acid sequence is amplified with polymerases. One particularly preferred method using polymerase-driven amplification is the polymerase chain reaction (PCR). The polymerase chain reaction and other polymerase-driven amplification assays can achieve over a million-fold increase in

copy number through the use of polymerase-driven amplification cycles. Once amplified, the resulting nucleic acid can be sequenced or used as a substrate for DNA probes.

When the probes are used to detect the presence of the target sequences (for example, in screening for cancer susceptibility), the biological sample to be analyzed, such as blood or serum, may be treated, if desired, to extract the nucleic acids. The sample nucleic acid may be prepared in various ways to facilitate detection of the target sequence; e.g. denaturation, restriction digestion, electrophoresis or dot blotting. The targeted region of the analyte nucleic acid usually must be at least partially single-stranded to form hybrids with the targeting sequence of the probe. If the sequence is naturally single-stranded, denaturation will not be required. However, if the sequence is double-stranded, the sequence will probably need to be denatured. Denaturation can be carried out by various techniques known in the art.

Analyte nucleic acid and probe are incubated under conditions which promote stable hybrid formation of the target sequence in the probe with the putative targeted sequence in the analyte. The region of the probes which is used to bind to the analyte can be made completely complementary to the targeted region of human chromosome 17. Therefore, high stringency conditions are desirable in order to prevent false positives. However, conditions of high stringency are used only if the probes are complementary to regions of the chromosome which are unique in the genome. The stringency of hybridization is determined by a number of factors during hybridization and during the washing procedure, including temperature, ionic strength, base composition, probe length, and concentration of formamide. These factors are outlined in, for example, Maniatis *et al.*, 1982 and Sambrook *et al.*, 1989. Under certain circumstances, the formation of higher order hybrids, such as triplexes, quadraplexes, etc., may be desired to provide the means of detecting target sequences.

Detection, if any, of the resulting hybrid is usually accomplished by the use of labeled probes. Alternatively, the probe may be unlabeled, but may be detectable by specific binding with a ligand which is labeled, either directly or indirectly. Suitable labels, and methods for labeling probes and ligands are known in the art, and include, for example, radioactive labels which may be incorporated by known methods (e.g., nick translation, random priming or kinasing), biotin, fluorescent groups, chemiluminescent groups (e.g., dioxetanes, particularly triggered dioxetanes), enzymes, antibodies, gold nanoparticles and the like. Variations of this basic scheme are known in the art, and include those variations that facilitate separation of the hybrids to be detected from extraneous materials and/or that amplify the signal from the labeled moiety. A number of these variations are reviewed in, e.g., Matthews and Kricka, 1988;

Landegren *et al.*, 1988; Mifflin, 1989; U.S. Patent 4,868,105, and in EPO Publication No. 225,807.

As noted above, non-PCR based screening assays are also contemplated in this invention. This procedure hybridizes a nucleic acid probe (or an analog such as a methyl phosphonate backbone replacing the normal phosphodiester), to the low level DNA target. This probe may have an enzyme covalently linked to the probe, such that the covalent linkage does not interfere with the specificity of the hybridization. This enzyme-probe-conjugate-target nucleic acid complex can then be isolated away from the free probe enzyme conjugate and a substrate is added for enzyme detection. Enzymatic activity is observed as a change in color development or luminescent output resulting in a 10^3-10^6 increase in sensitivity. For an example relating to the preparation of oligodeoxynucleotide-alkaline phosphatase conjugates and their use as hybridization probes see Jablonski *et al.*, 1986.

Two-step label amplification methodologies are known in the art. These assays work on the principle that a small ligand (such as digoxigenin, biotin, or the like) is attached to a nucleic acid probe capable of specifically binding HPC2. Allele specific probes are also contemplated within the scope of this example and exemplary allele specific probes include probes encompassing the predisposing or potentially predisposing mutations summarized in Table 9 of this patent application.

In one example, the small ligand attached to the nucleic acid probe is specifically recognized by an antibody-enzyme conjugate. In one embodiment of this example, digoxigenin is attached to the nucleic acid probe. Hybridization is detected by an antibody-alkaline phosphatase conjugate which turns over a chemiluminescent substrate. For methods for labeling nucleic acid probes according to this embodiment see Martin *et al.*, 1990. In a second example, the small ligand is recognized by a second ligand-enzyme conjugate that is capable of specifically complexing to the first ligand. A well known embodiment of this example is the biotin-avidin type of interactions. For methods for labeling nucleic acid probes and their use in biotin-avidin based assays see Rigby *et al.*, 1977 and Nguyen *et al.*, 1992.

It is also contemplated within the scope of this invention that the nucleic acid probe assays of this invention will employ a cocktail of nucleic acid probes capable of detecting HPC2. Thus, in one example to detect the presence of HPC2 in a cell sample, more than one probe complementary to HPC2 is employed and in particular the number of different probes is alternatively 2, 3, or 5 different nucleic acid probe sequences. In another example, to detect the presence of mutations in the HPC2 gene sequence in a patient, more than one probe

complementary to HPC2 is employed where the cocktail includes probes capable of binding to the allele-specific mutations identified in populations of patients with alterations in HPC2. In this embodiment, any number of probes can be used, and will preferably include probes corresponding to the major gene mutations identified as predisposing an individual to prostate cancer.

Methods of Use: Peptide Diagnosis and Diagnostic Kits

The neoplastic condition of lesions can also be detected on the basis of the alteration of wild-type HPC2 polypeptide. Such alterations can be determined by sequence analysis in accordance with conventional techniques. More preferably, antibodies (polyclonal or monoclonal) are used to detect differences in, or the absence of, HPC2 peptides. The antibodies may be prepared as discussed above under the heading "Antibodies" and as further shown in Examples 16 and 17. Other techniques for raising and purifying antibodies are well known in the art and any such techniques may be chosen to achieve the preparations claimed in this invention. In a preferred embodiment of the invention, antibodies will immunoprecipitate HPC2 proteins from solution as well as react with HPC2 protein on Western or immunoblots of polyacrylamide gels. In another preferred embodiment, antibodies will detect HPC2 proteins in paraffin or frozen tissue sections, using immunocytochemical techniques.

Preferred embodiments relating to methods for detecting HPC2 or its mutations include enzyme linked immunosorbent assays (ELISA), radioimmunoassays (RIA), immunoradiometric assays (IRMA) and immunoenzymatic assays (IEMA), including sandwich assays using monoclonal and/or polyclonal antibodies. Exemplary sandwich assays are described by David *et al.* in U.S. Patent Nos. 4,376,110 and 4,486,530, hereby incorporated by reference, and exemplified in Example 19.

Methods of Use: Drug Screening

This invention is particularly useful for screening compounds by using the HPC2, ELAC1 or ELAC2 polypeptide or binding fragment thereof in any of a variety of drug screening techniques.

The HPC2, ELAC1 or ELAC2 polypeptide or fragment employed in such a test may either be free in solution, affixed to a solid support, or borne on a cell surface. One method of drug screening utilizes eucaryotic or procaryotic host cells which are stably transformed with recombinant polynucleotides expressing the polypeptide or fragment, preferably in competitive

binding assays. Such cells, either in viable or fixed form, can be used for standard binding assays. One may measure, for example, for the formation of complexes between an HPC2, ELAC1 or ELAC2 polypeptide or fragment and the agent being tested, or examine the degree to which the formation of a complex between an HPC2, ELAC1 or ELAC2 polypeptide or fragment and a known ligand is interfered with by the agent being tested.

Thus, the present invention provides methods of screening for drugs comprising contacting such an agent with an HPC2, ELAC1 or ELAC2 polypeptide or fragment thereof and assaying (i) for the presence of a complex between the agent and the HPC2, ELAC1 or ELAC2 polypeptide or fragment, or (ii) for the presence of a complex between the HPC2, ELAC1 or ELAC2 polypeptide or fragment and a ligand, by methods well known in the art. In such competitive binding assays the HPC2, ELAC1 or ELAC2 polypeptide or fragment is typically labeled. Free HPC2, ELAC1 or ELAC2 polypeptide or fragment is separated from that present in a protein:protein complex, and the amount of free (i.e., uncomplexed) label is a measure of the binding of the agent being tested to HPC2, ELAC1 or ELAC2 or its interference with HPC2:ligand, ELAC1:ligand or ELAC2:ligand binding, respectively. One may also measure the amount of bound, rather than free, HPC2, ELAC1 or ELAC2. It is also possible to label the ligand rather than the HPC2, ELAC1 or ELAC2 and to measure the amount of ligand binding to HPC2, ELAC1 or ELAC2 in the presence and in the absence of the drug being tested.

Another technique for drug screening provides high throughput screening for compounds having suitable binding affinity to the HPC2, ELAC1 or ELAC2 polypeptides and is described in detail in Geysen (published PCT WO 84/03564). Briefly stated, large numbers of different small peptide test compounds are synthesized on a solid substrate, such as plastic pins or some other surface. The peptide test compounds are reacted with HPC2, ELAC1 or ELAC2 polypeptide and washed. Bound HPC2, ELAC1 or ELAC2 polypeptide is then detected by methods well known in the art.

Purified HPC2, ELAC1 or ELAC2 can be coated directly onto plates for use in the aforementioned drug screening techniques. However, non-neutralizing antibodies to the polypeptide can be used to capture antibodies to immobilize the HPC2, ELAC1 or ELAC2 polypeptide on the solid phase.

This invention also contemplates the use of competitive drug screening assays in which neutralizing antibodies capable of specifically binding the HPC2, ELAC1 or ELAC2 polypeptide compete with a test compound for binding to the HPC2, ELAC1 or ELAC2 polypeptide or fragments thereof. In this manner, the antibodies can be used to detect the

presence of any peptide which shares one or more antigenic determinants of the HPC2, ELAC1 or ELAC2 polypeptide.

A further technique for drug screening involves the use of host eukaryotic cell lines or cells (such as described above) which have a nonfunctional HPC2, ELAC1 or ELAC2 gene. These host cell lines or cells are defective at the HPC2, ELAC1 or ELAC2 polypeptide level. The host cell lines or cells are grown in the presence of drug compound. The rate of growth of the host cells is measured to determine if the compound is capable of regulating the growth of HPC2, ELAC1 or ELAC2 defective cells.

Briefly, a method of screening for a substance which modulates activity of a polypeptide may include contacting one or more test substances with the polypeptide in a suitable reaction medium, testing the activity of the treated polypeptide and comparing that activity with the activity of the polypeptide in comparable reaction medium untreated with the test substance or substances. A difference in activity between the treated and untreated polypeptides is indicative of a modulating effect of the relevant test substance or substances.

Prior to or as well as being screened for modulation of activity, test substances may be screened for ability to interact with the polypeptide, e.g., in a yeast two-hybrid system (e.g., Bartel et al., 1993; Fields and Song, 1989; Chevray and Nathans, 1992; Lee et al., 1995). This system may be used as a coarse screen prior to testing a substance for actual ability to modulate activity of the polypeptide. Alternatively, the screen could be used to screen test substances for binding to an HPC2, ELAC1 or ELAC2 specific binding partner, or to find mimetics of an HPC2, ELAC1 or ELAC2 polypeptide.

Methods of Use: Rational Drug Design

The goal of rational drug design is to produce structural analogs of biologically active polypeptides of interest or of small molecules with which they interact (e.g., agonists, antagonists, inhibitors) in order to fashion drugs which are, for example, more active or stable forms of the polypeptide, or which, e.g., enhance or interfere with the function of a polypeptide *in vivo*. See, e.g., Hodgson, 1991. In one approach, one first determines the three-dimensional structure of a protein of interest (e.g., HPC2 polypeptide) or, for example, of the HPC2-receptor or ligand complex, by x-ray crystallography, by computer modeling or most typically, by a combination of approaches. Less often, useful information regarding the structure of a polypeptide may be gained by modeling based on the structure of homologous proteins. An example of rational drug design is the development of HIV protease inhibitors (Erickson *et al.*,

1990). In addition, peptides (e.g., HPC2 polypeptide) are analyzed by an alanine scan (Wells, 1991). In this technique, an amino acid residue is replaced by Ala, and its effect on the peptide's activity is determined. Each of the amino acid residues of the peptide is analyzed in this manner to determine the important regions of the peptide.

It is also possible to isolate a target-specific antibody, selected by a functional assay, and then to solve its crystal structure. In principle, this approach yields a pharmacore upon which subsequent drug design can be based. It is possible to bypass protein crystallography altogether by generating anti-idiotypic antibodies (anti-ids) to a functional, pharmacologically active antibody. As a mirror image of a mirror image, the binding site of the anti-ids would be expected to be an analog of the original receptor. The anti-id could then be used to identify and isolate peptides from banks of chemically or biologically produced banks of peptides. Selected peptides would then act as the pharmacore.

Thus, one may design drugs which have, e.g., improved HPC2, ELAC1 or ELAC2 polypeptide activity or stability or which act as inhibitors, agonists, antagonists, etc. of HPC2, ELAC1 or ELAC2 polypeptide activity. By virtue of the availability of cloned HPC2, ELAC1 and ELAC2 sequences, sufficient amounts of the HPC2, ELAC1 or ELAC2 polypeptide may be made available to perform such analytical studies as x-ray crystallography. In addition, the knowledge of the HPC2, ELAC1 and ELAC2 protein sequences provided herein will guide those employing computer modeling techniques in place of, or in addition to x-ray crystallography.

Following identification of a substance which modulates or affects polypeptide activity, the substance may be investigated further. Furthermore, it may be manufactured and/or used in preparation, i.e., manufacture or formulation, or a composition such as a medicament, pharmaceutical composition or drug. These may be administered to individuals.

Thus, the present invention extends in various aspects not only to a substance identified using a nucleic acid molecule as a modulator of polypeptide activity, in accordance with what is disclosed herein, but also a pharmaceutical composition, medicament, drug or other composition comprising such a substance, a method comprising administration of such a composition comprising such a substance, a method comprising administration of such a composition to a patient, e.g., for treatment of prostate cancer, use of such a substance in the manufacture of a composition for administration, e.g., for treatment of prostate cancer, and a method of making a pharmaceutical composition comprising admixing such a substance with a pharmaceutically acceptable excipient, vehicle or carrier, and optionally other ingredients.

A substance identified as a modulator of polypeptide function may be peptide or non-peptide in nature. Non-peptide "small molecules" are often preferred for many *in vivo* pharmaceutical uses. Accordingly, a mimetic or mimic of the substance (particularly if a peptide) may be designed for pharmaceutical use.

The designing of mimetics to a known pharmaceutically active compound is a known approach to the development of pharmaceuticals based on a "lead" compound. This might be desirable where the active compound is difficult or expensive to synthesize or where it is unsuitable for a particular method of administration, e.g., pure peptides are unsuitable active agents for oral compositions as they tend to be quickly degraded by proteases in the alimentary canal. Mimetic design, synthesis and testing is generally used to avoid randomly screening large numbers of molecules for a target property.

There are several steps commonly taken in the design of a mimetic from a compound having a given target property. First, the particular parts of the compound that are critical and/or important in determining the target property are determined. In the case of a peptide, this can be done by systematically varying the amino acid residues in the peptide, e.g., by substituting each residue in turn. Alanine scans of peptide are commonly used to refine such peptide motifs. These parts or residues constituting the active region of the compound are known as its "pharmacophore".

Once the pharmacophore has been found, its structure is modeled according to its physical properties, e.g., stereochemistry, bonding, size and/or charge, using data from a range of sources, e.g., spectroscopic techniques, x-ray diffraction data and NMR. Computational analysis, similarity mapping (which models the charge and/or volume of a pharmacophore, rather than the bonding between atoms) and other techniques can be used in this modeling process.

In a variant of this approach, the three-dimensional structure of the ligand and its binding partner are modeled. This can be especially useful where the ligand and/or binding partner change conformation on binding, allowing the model to take account of this in the design of the mimetic.

A template molecule is then selected onto which chemical groups which mimic the pharmacophore can be grafted. The template molecule and the chemical groups grafted onto it can conveniently be selected so that the mimetic is easy to synthesize, is likely to be pharmacologically acceptable, and does not degrade *in vivo*, while retaining the biological activity of the lead compound. Alternatively, where the mimetic is peptide-based, further

stability can be achieved by cyclizing the peptide, increasing its rigidity. The mimetic or mimetics found by this approach can then be screened to see whether they have the target property, or to what extent they exhibit it. Further optimization or modification can then be carried out to arrive at one or more final mimetics for *in vivo* or clinical testing.

Methods of Use: Gene Therapy

According to the present invention, a method is also provided of supplying wild-type HPC2 function to a cell which carries mutant HPC2 alleles. Supplying such a function should suppress neoplastic growth of the recipient cells. The wild-type HPC2 gene or a part of the gene may be introduced into the cell in a vector such that the gene remains extrachromosomal. In such a situation, the gene will be expressed by the cell from the extrachromosomal location. If a gene fragment is introduced and expressed in a cell carrying a mutant HPC2 allele, the gene fragment should encode a part of the HPC2 protein which is required for non-neoplastic growth of the cell. More preferred is the situation where the wild-type HPC2 gene or a part thereof is introduced into the mutant cell in such a way that it recombines with the endogenous mutant HPC2 gene present in the cell. Such recombination requires a double recombination event which results in the correction of the HPC2 gene mutation. Vectors for introduction of genes both for recombination and for extrachromosomal maintenance are known in the art, and any suitable vector may be used. Methods for introducing DNA into cells such as electroporation, calcium phosphate coprecipitation and viral transduction are known in the art, and the choice of method is within the competence of the practitioner. Cells transformed with the wild-type HPC2 gene can be used as model systems to study cancer remission and drug treatments which promote such remission.

As generally discussed above, the HPC2 gene or fragment, where applicable, may be employed in gene therapy methods in order to increase the amount of the expression products of such genes in cancer cells. Such gene therapy is particularly appropriate for use in both cancerous and pre-cancerous cells, in which the level of HPC2 polypeptide is absent or diminished compared to normal cells. It may also be useful to increase the level of expression of a given HPC2 gene even in those tumor cells in which the mutant gene is expressed at a "normal" level, but the gene product is not fully functional.

Gene therapy would be carried out according to generally accepted methods, for example, as described by Friedman (1991) or Culver (1996). Cells from a patient's tumor would be first analyzed by the diagnostic methods described above, to ascertain the production of

HPC2 polypeptide in the tumor cells. A virus or plasmid vector (see further details below), containing a copy of the HPC2 gene linked to expression control elements and capable of replicating inside the tumor cells, is prepared. Alternatively, the vector may be replication deficient and is replicated in helper cells for use in gene therapy. Suitable vectors are known, such as disclosed in U.S. Patent 5,252,479 and PCT published application WO 93/07282 and U.S. Patent Nos. 5,691,198; 5,747,469; 5,436,146 and 5,753,500. The vector is then injected into the patient, either locally at the site of the tumor or systemically (in order to reach any tumor cells that may have metastasized to other sites). If the transfected gene is not permanently incorporated into the genome of each of the targeted tumor cells, the treatment may have to be repeated periodically.

Gene transfer systems known in the art may be useful in the practice of the gene therapy methods of the present invention. These include viral and nonviral transfer methods. A number of viruses have been used as gene transfer vectors, including papovaviruses, e.g., SV40 (Madzak *et al.*, 1992), adenovirus (Berkner, 1992; Berkner *et al.*, 1988; Gorziglia and Kapikian, 1992; Quantin *et al.*, 1992; Rosenfeld *et al.*, 1992; Wilkinson and Akrigg, 1992; Stratford-Perricaudet *et al.*, 1990; Schneider *et al.*, 1998), vaccinia virus (Moss, 1992; Moss, 1996), adeno-associated virus (Muzyczka, 1992; Ohi *et al.*, 1990; Russell and Hirata, 1998), herpes viruses including HSV and EBV (Margolskee, 1992; Johnson *et al.*, 1992; Fink *et al.*, 1992; Breakefield and Geller, 1987; Freese *et al.*, 1990; Fink et al., 1996), lentiviruses (Naldini et al., 1996), Sindbis and Semliki Forest virus (Berglund et al., 1993), and retroviruses of avian (Bandyopadhyay and Temin, 1984; Petropoulos *et al.*, 1992), murine (Miller, 1992; Miller *et al.*, 1985; Sorge *et al.*, 1984; Mann and Baltimore, 1985; Miller *et al.*, 1988), and human origin (Shimada *et al.*, 1991; Helseth *et al.*, 1990; Page *et al.*, 1990; Buchschacher and Panganiban, 1992). Most human gene therapy protocols have been based on disabled murine retroviruses, although adenovirus and adeno-associated virus are also being used.

Nonviral gene transfer methods known in the art include chemical techniques such as calcium phosphate coprecipitation (Graham and van der Eb, 1973; Pellicer *et al.*, 1980); mechanical techniques, for example microinjection (Anderson *et al.*, 1980; Gordon *et al.*, 1980; Brinster *et al.*, 1981; Costantini and Lacy, 1981); membrane fusion-mediated transfer via liposomes (Felgner *et al.*, 1987; Wang and Huang, 1989; Kaneda et al, 1989; Stewart *et al.*, 1992; Nabel *et al.*, 1990; Lim *et al.*, 1991); and direct DNA uptake and receptor-mediated DNA transfer (Wolff *et al.*, 1990; Wu *et al.*, 1991; Zenke *et al.*, 1990; Wu *et al.*, 1989; Wolff *et al.*, 1991; Wagner *et al.*, 1990; Wagner *et al.*, 1991; Cotten *et al.*, 1990; Curiel *et al.*, 1991; Curiel *et*

al., 1992). Viral-mediated gene transfer can be combined with direct *in vivo* gene transfer using liposome delivery, allowing one to direct the viral vectors to the tumor cells and not into the surrounding nondividing cells. Alternatively, the retroviral vector producer cell line can be injected into tumors (Culver *et al.*, 1992). Injection of producer cells would then provide a continuous source of vector particles. This technique has been approved for use in humans with inoperable brain tumors.

In an approach which combines biological and physical gene transfer methods, plasmid DNA of any size is combined with a polylysine-conjugated antibody specific to the adenovirus hexon protein, and the resulting complex is bound to an adenovirus vector. The trimolecular complex is then used to infect cells. The adenovirus vector permits efficient binding, internalization, and degradation of the endosome before the coupled DNA is damaged. For other techniques for the delivery of adenovirus based vectors see Schneider et al. (1998) and U.S. Patent Nos. 5,691,198; 5,747,469; 5,436,146 and 5,753,500.

Liposome/DNA complexes have been shown to be capable of mediating direct *in vivo* gene transfer. While in standard liposome preparations the gene transfer process is nonspecific, localized *in vivo* uptake and expression have been reported in tumor deposits, for example, following direct *in situ* administration (Nabel, 1992).

Expression vectors in the context of gene therapy are meant to include those constructs containing sequences sufficient to express a polynucleotide that has been cloned therein. In viral expression vectors, the construct contains viral sequences sufficient to support packaging of the construct. If the polynucleotide encodes *HPC2*, expression will produce HPC2. If the polynucleotide encodes an antisense polynucleotide or a ribozyme, expression will produce the antisense polynucleotide or ribozyme. Thus in this context, expression does not require that a protein product be synthesized. In addition to the polynucleotide cloned into the expression vector, the vector also contains a promoter functional in eukaryotic cells. The cloned polynucleotide sequence is under control of this promoter. Suitable eukaryotic promoters include those described above. The expression vector may also include sequences, such as selectable markers and other sequences described herein.

Gene transfer techniques which target DNA directly to prostate tissues, e.g., epithelial cells of the prostate, are preferred. Receptor-mediated gene transfer, for example, is accomplished by the conjugation of DNA (usually in the form of covalently closed supercoiled plasmid) to a protein ligand via polylysine. Ligands are chosen on the basis of the presence of the corresponding ligand receptors on the cell surface of the target cell/tissue type. One

appropriate receptor/ligand pair may include the estrogen receptor and its ligand, estrogen (and estrogen analogues). These ligand-DNA conjugates can be injected directly into the blood if desired and are directed to the target tissue where receptor binding and internalization of the DNA-protein complex occurs. To overcome the problem of intracellular destruction of DNA, coinfection with adenovirus can be included to disrupt endosome function.

The therapy involves two steps which can be performed singly or jointly. In the first step, prepubescent females who carry an HPC2 susceptibility allele are treated with a gene delivery vehicle such that some or all of their mammary ductal epithelial precursor cells receive at least one additional copy of a functional normal HPC2 allele. In this step, the treated individuals have reduced risk of prostate cancer to the extent that the effect of the susceptible allele has been countered by the presence of the normal allele. In the second step of a preventive therapy, predisposed young females, in particular women who have received the proposed gene therapeutic treatment, undergo hormonal therapy to mimic the effects on the prostate of a full term pregnancy.

Methods of Use: Peptide Therapy

Peptides which have HPC2, ELAC1 or ELAC2 activity can be supplied to cells which carry mutant or missing HPC2, ELAC1 or ELAC2 alleles. Protein can be produced by expression of the cDNA sequence in bacteria, for example, using known expression vectors. Alternatively, HPC2, ELAC1 or ELAC2 polypeptide can be extracted from HPC2-, ELAC1- or ELAC2-producing mammalian cells. In addition, the techniques of synthetic chemistry can be employed to synthesize HPC2, ELAC1 or ELAC2 protein. Any of such techniques can provide the preparation of the present invention which comprises the HPC2, ELAC1 or ELAC2 protein. Preparation is substantially free of other human proteins. This is most readily accomplished by synthesis in a microorganism or *in vitro*.

Active HPC2, ELAC1 or ELAC2 molecules can be introduced into cells by microinjection or by use of liposomes, for example. Alternatively, some active molecules may be taken up by cells, actively or by diffusion. Extracellular application of the HPC2, ELAC1 or ELAC2 gene product may be sufficient to affect tumor growth. Supply of molecules with HPC2 activity should lead to partial reversal of the neoplastic state. Other molecules with HPC2 activity (for example, peptides, drugs or organic compounds) may also be used to effect such a reversal. Modified polypeptides having substantially similar function are also used for peptide therapy.

Methods of Use: Transformed Hosts

Similarly, cells and animals which carry a mutant HPC2, ELAC1 or ELAC2 allele can be used as model systems to study and test for substances which have potential as therapeutic agents. The cells are typically cultured epithelial cells. These may be isolated from individuals with HPC2, ELAC1 or ELAC2 mutations, either somatic or germline. Alternatively, the cell line can be engineered to carry the mutation in the HPC2, ELAC1 or ELAC2 allele, as described above. After a test substance is applied to the cells, the neoplastically transformed phenotype of the cell is determined. Any trait of neoplastically transformed cells can be assessed, including anchorage-independent growth, tumorigenicity in nude mice, invasiveness of cells, and growth factor dependence. Assays for each of these traits are known in the art.

Animals for testing therapeutic agents can be selected after mutagenesis of whole animals or after treatment of germline cells or zygotes. Such treatments include insertion of mutant HPC2, ELAC1 or ELAC2 alleles, usually from a second animal species, as well as insertion of disrupted homologous genes. Alternatively, the endogenous HPC2, ELAC1 or ELAC2 gene(s) of the animals may be disrupted by insertion or deletion mutation or other genetic alterations using conventional techniques (Capecchi, 1989; Valancius and Smithies, 1991; Hasty *et al.*, 1991; Shinkai *et al.*, 1992; Mombaerts *et al.*, 1992; Philpott *et al.*, 1992; Snouwaert *et al.*, 1992; Donehower *et al.*, 1992) to produce knockout or transplacement animals. A transplacement is similar to a knockout because the endogenous gene is replaced, but in the case of a transplacement the replacement is by another version of the same gene. After test substances have been administered to the animals, the growth of tumors must be assessed. If the test substance prevents or suppresses the growth of tumors, then the test substance is a candidate therapeutic agent for the treatment of the cancers identified herein. These animal models provide an extremely important testing vehicle for potential therapeutic products.

In one embodiment of the invention, transgenic animals are produced which contain a functional transgene encoding a functional HPC2, ELAC1 or ELAC2 polypeptide or variants thereof. Transgenic animals expressing *HPC2*, *ELAC1* or *ELAC2* transgenes, recombinant cell lines derived from such animals and transgenic embryos may be useful in methods for screening for and identifying agents that induce or repress function of HPC2, ELAC1 or ELAC2. Transgenic animals of the present invention also can be used as models for studying indications such as disease.

In one embodiment of the invention, an *HPC2*, *ELAC1* or *ELAC2* transgene is introduced into a non-human host to produce a transgenic animal expressing a human or murine *HPC2*, *ELAC1* or *ELAC2* gene. The transgenic animal is produced by the integration of the transgene into the genome in a manner that permits the expression of the transgene. Methods for producing transgenic animals are generally described by Wagner and Hoppe (U.S. Patent No. 4,873,191; which is incorporated herein by reference), Brinster *et al.* 1985; which is incorporated herein by reference in its entirety) and in "Manipulating the Mouse Embryo; A Laboratory Manual" 2nd edition (eds., Hogan, Beddington, Costantini and Long, Cold Spring Harbor Laboratory Press, 1994; which is incorporated herein by reference in its entirety).

It may be desirable to replace the endogenous *HPC2*, *ELAC1* or *ELAC2* by homologous recombination between the transgene and the endogenous gene; or the endogenous gene may be eliminated by deletion as in the preparation of "knock-out" animals. Typically, an *HPC2*, *ELAC1* or *ELAC2* gene flanked by genomic sequences is transferred by microinjection into a fertilized egg. The microinjected eggs are implanted into a host female, and the progeny are screened for the expression of the transgene. Transgenic animals may be produced from the fertilized eggs from a number of animals including, but not limited to reptiles, amphibians, birds, mammals, and fish. Within a particularly preferred embodiment, transgenic mice are generated which overexpress HPC2 or express a mutant form of the polypeptide. Alternatively, the absence of an *HPC2*, *ELAC1* or *ELAC2* in "knock-out" mice permits the study of the effects that loss of HPC2, ELAC1 or ELAC2 protein has on a cell *in vivo*. Knock-out mice also provide a model for the development of HPC2-related cancers.

Methods for producing knockout animals are generally described by Shastry (1995, 1998) and Osterrieder and Wolf (1998). The production of conditional knockout animals, in which the gene is active until knocked out at the desired time is generally described by Feil et al. (1996), Gagneten et al. (1997) and Lobe and Nagy (1998). Each of these references is incorporated herein by reference.

As noted above, transgenic animals and cell lines derived from such animals may find use in certain testing experiments. In this regard, transgenic animals and cell lines capable of expressing wild-type or mutant HPC2, ELAC1 or ELAC2 may be exposed to test substances. These test substances can be screened for the ability to reduce overexpression of wild-type *HPC2*, *ELAC1* or *ELAC2* or impair the expression or function of mutant HPC2, ELAC1 or ELAC2.

Pharmaceutical Compositions and Routes of Administration

The HPC2, ELAC1 or ELAC2 polypeptides, antibodies, peptides and nucleic acids of the present invention can be formulated in pharmaceutical compositions, which are prepared according to conventional pharmaceutical compounding techniques. See, for example, Remington's Pharmaceutical Sciences, 18th Ed. (1990, Mack Publishing Co., Easton, PA). The composition may contain the active agent or pharmaceutically acceptable salts of the active agent. These compositions may comprise, in addition to one of the active substances, a pharmaceutically acceptable excipient, carrier, buffer, stabilizer or other materials well known in the art. Such materials should be non-toxic and should not interfere with the efficacy of the active ingredient. The carrier may take a wide variety of forms depending on the form of preparation desired for administration, e.g., intravenous, oral, intrathecal, epineural or parenteral.

For oral administration, the compounds can be formulated into solid or liquid preparations such as capsules, pills, tablets, lozenges, melts, powders, suspensions or emulsions. In preparing the compositions in oral dosage form, any of the usual pharmaceutical media may be employed, such as, for example, water, glycols, oils, alcohols, flavoring agents, preservatives, coloring agents, suspending agents, and the like in the case of oral liquid preparations (such as, for example, suspensions, elixirs and solutions); or carriers such as starches, sugars, diluents, granulating agents, lubricants, binders, disintegrating agents and the like in the case of oral solid preparations (such as, for example, powders, capsules and tablets). Because of their ease in administration, tablets and capsules represent the most advantageous oral dosage unit form, in which case solid pharmaceutical carriers are obviously employed. If desired, tablets may be sugar-coated or enteric-coated by standard techniques. The active agent can be encapsulated to make it stable to passage through the gastrointestinal tract while at the same time allowing for passage across the blood brain barrier. See for example, WO 96/11698.

For parenteral administration, the compound may be dissolved in a pharmaceutical carrier and administered as either a solution or a suspension. Illustrative of suitable carriers are water, saline, dextrose solutions, fructose solutions, ethanol, or oils of animal, vegetative or synthetic origin. The carrier may also contain other ingredients, for example, preservatives, suspending agents, solubilizing agents, buffers and the like. When the compounds are being administered intrathecally, they may also be dissolved in cerebrospinal fluid.

The active agent is preferably administered in a therapeutically effective amount. The actual amount administered, and the rate and time-course of administration, will depend on the nature and severity of the condition being treated. Prescription of treatment, e.g. decisions on

dosage, timing, etc., is within the responsibility of general practitioners or specialists, and typically takes account of the disorder to be treated, the condition of the individual patient, the site of delivery, the method of administration and other factors known to practitioners. Examples of techniques and protocols can be found in *Remington's Pharmaceutical Sciences*.

Alternatively, targeting therapies may be used to deliver the active agent more specifically to certain types of cell, by the use of targeting systems such as antibodies or cell specific ligands. Targeting may be desirable for a variety of reasons, e.g. if the agent is unacceptably toxic, or if it would otherwise require too high a dosage, or if it would not otherwise be able to enter the target cells.

Instead of administering these agents directly, they could be produced in the target cell, e.g. in a viral vector such as described above or in a cell based delivery system such as described in U.S. Patent No. 5,550,050 and published PCT application Nos. WO 92/19195, WO 94/25503, WO 95/01203, WO 95/05452, WO 96/02286, WO 96/02646, WO 96/40871, WO 96/40959 and WO 97/12635, designed for implantation in a patient. The vector could be targeted to the specific cells to be treated, or it could contain regulatory elements which are more tissue specific to the target cells. The cell based delivery system is designed to be implanted in a patient's body at the desired target site and contains a coding sequence for the active agent. Alternatively, the agent could be administered in a precursor form for conversion to the active form by an activating agent produced in, or targeted to, the cells to be treated. See for example, EP 425,731A and WO 90/07936.

As disclosed in the following Examples, on the basis of segregating mutations of *HPC2* in kindreds 4102 and 4289, plus association between carriage of the common missense changes Leu 217 and Thr 541 with a diagnosis of prostate cancer, we conclude that *HPC2* is a prostate cancer susceptibility gene.

While a 1641 insG frameshift found in kindred 4102 will clearly disrupt protein function, this is not obviously the case for the His 781 missense change in kindred 4289. Interestingly, this missense change occurred on a chromosome that also carries Leu 217 and Thr 541. Thus one might entertain an additive hypothesis to explain the relative strength of the three missense bearing alleles that we have observed. Substitution of Leu for Ser 217 may change the character of a normally hydrophilic segment of the protein; the phenotype conferred is sufficiently modest that it is only detected when the variant is homozygous. Ala 541 is immediately adjacent to the histidine motif. At the position corresponding to Ala 541 in the ELAC1/2, CPSF73 and PSO2

gene families, the most common residue is alanine; when not alanine, the residue is hydrophobic, amide, or basic (Figures 6A-B, 9 and12). Although threonine is observed at this position in other histidine motif containing gene families, it is rare or absent in these three closely related gene families. Thus, from sequence conservation considerations, it is quite reasonable that the Leu 217 + Thr 541 allele should be more deleterious than Leu 217 alone, apparently sufficiently deleterious to be detected in a co-dominant to dominant association test. The kindred 4289 allele carries all three missense changes, Leu 217, Thr 541 and His 781. Examination of the pedigree suggests that the allele is dominant and sufficiently deleterious to demonstrate visible segregation with prostate cancer in an extended pedigree. Interestingly, the youngest affected carrier of this variant, 4289.003, is homozygous for Leu 217 and Thr 541. Thus his mother, the second ovarian cancer case in the pedigree, is an obligate carrier of a Leu 217 + Thr 541 allele. The observation of two ovarian cancer cases in this pedigree, both of whom carry deleterious alleles of *ELAC2*, is consistent with the possibility that the phenotype conferred by deleterious variants in this gene is not restricted to prostate cancer susceptibility.

The potential contributions of the androgen receptor CAG repeat and *SRD5A2* Ala 49 Thr missense change to prostate cancer risk were first detected in association studies using sporadic cases and unaffected controls. However, straightforward deduction from the considerable literature on sib pair analyses would predict that such sequence variants should be enriched among affected sibs versus isolated cases, and it follows that such sequence variants should contribute to a larger fraction of familial than truly sporadic prostate cancer cases. Thus one might expect genotypes at moderate risk susceptibility genes such as the androgen receptor, *SRD5A2*, and the common missense changes in *HPC2/ELAC2*, to confound linkage studies aimed at detecting and localizing lower prevalence, higher risk susceptibility genes. However, inclusion of genotype information from pedigree members at multiple moderate risk loci may allow refined definition of the liability classes used by multipoint linkage software, thereby increasing the power of the analysis. Stratification of cases by genotype would also facilitate positional cloning projects by providing another criterion by which to distinguish between true recombinant carriers and confounding sporadic cases.

The genetic data presented demonstrate that there are deleterious sequence variants in *HPC2/ELAC2* that contribute to prostate cancer risk. Elucidating the functional alteration by which a moderate risk sequence variant such as Leu 217 contributes to a late onset pathology could prove difficult because its manifestation could be quite subtle. However, a mutation as dramatic as a frameshift leading to protein truncation within the likely active site of an enzyme

should have a more easily detected effect on cell physiology. Conservation of the C-terminal domain of the gene through the eubacteria and archaebacteria, combined with the observation that the *S. cerevisiae* ortholog YRK079C is essential, emphasize that the function of the ELAC1/2 gene family is of fundamental biological interest.

The identification of the association between the HPC2 gene mutations and prostate cancer permits the early presymptomatic screening of individuals to identify those at risk for developing prostate cancer. To identify such individuals, HPC2 alleles are screened for mutations either directly or after cloning the alleles. The alleles are tested for the presence of nucleic acid sequence differences from the normal allele using any suitable technique, including but not limited to, one of the following methods: fluorescent *in situ* hybridization (FISH), direct DNA sequencing, PFGE analysis, Southern blot analysis, single stranded conformation analysis (SSCP), linkage analysis, RNase protection assay, allele specific oligonucleotide (ASO), dot blot analysis and PCR-SSCP analysis. Also useful is the recently developed technique of DNA microchip technology. For example, either (1) the nucleotide sequence of both the cloned alleles and normal HPC2 gene or appropriate fragment (coding sequence or genomic sequence) are determined and then compared, or (2) the RNA transcripts of the HPC2 gene or gene fragment are hybridized to single stranded whole genomic DNA from an individual to be tested, and the resulting heteroduplex is treated with Ribonuclease A (RNase A) and run on a denaturing gel to detect the location of any mismatches. Two of these methods can be carried out according to the following procedures.

The alleles of the *HPC2* gene in an individual to be tested are cloned using conventional techniques. For example, a blood sample is obtained from the individual. The genomic DNA isolated from the cells in this sample is partially digested to an average fragment size of approximately 20 kb. Fragments in the range from 18-21 kb are isolated. The resulting fragments are ligated into an appropriate vector. The sequences of the clones are then determined and compared to the normal *HPC2* gene.

Alternatively, polymerase chain reactions (PCRs) are performed with primer pairs for the 5 region or the exons of the *HPC2* gene. PCRs can also be performed with primer pairs based on any sequence of the normal *HPC2* gene. For example, primer pairs for one of the introns can be prepared and utilized. Finally, RT-PCR can also be performed on the mRNA. The amplified products are then analyzed by single stranded conformation polymorphisms (SSCP) using conventional techniques to identify any differences and these are then sequenced and compared to the normal gene sequence.

Individuals can be quickly screened for common *HPC2* gene variants by amplifying the individual s DNA using suitable primer pairs and analyzing the amplified product, e.g., by dot-blot hybridization using allele-specific oligonucleotide probes.

The second method employs RNase A to assist in the detection of differences between the normal *HPC2* gene and defective genes. This comparison is performed in steps using small (~500 bp) restriction fragments of the *HPC2* gene as the probe. First, the *HPC2* gene is digested with a restriction enzyme(s) that cuts the gene sequence into fragments of approximately 500 bp. These fragments are separated on an electrophoresis gel, purified from the gel and cloned individually, in both orientations, into an SP6 vector (e.g., pSP64 or pSP65). The SP6-based plasmids containing inserts of the *HPC2* gene fragments are transcribed *in vitro* using the SP6 transcription system, well known in the art, in the presence of [α-^{32}P]GTP, generating radiolabeled RNA transcripts of both strands of the gene.

Individually, these RNA transcripts are used to form heteroduplexes with the allelic DNA using conventional techniques. Mismatches that occur in the RNA:DNA heteroduplex, owing to sequence differences between the *HPC2* fragment and the *HPC2* allele subclone from the individual, result in cleavage in the RNA strand when treated with RNase A. Such mismatches can be the result of point mutations or small deletions in the individual's allele. Cleavage of the RNA strand yields two or more small RNA fragments, which run faster on the denaturing gel than the RNA probe itself.

Any differences which are found, will identify an individual as having a molecular variant of the *HPC2*. These variants can take a number of forms. The most severe forms would be frame shift mutations or large deletions which would cause the gene to code for an abnormal protein or one which would significantly alter protein expression. Less severe disruptive mutations would include small in-frame deletions and nonconservative base pair substitutions which would have a significant effect on the protein produced, such as changes to or from a cysteine residue, from a basic to an acidic amino acid or vice versa, from a hydrophobic to hydrophilic amino acid or vice versa, or other mutations which would affect secondary or tertiary protein structure. Silent mutations or those resulting in conservative amino acid substitutions would not generally be expected to disrupt protein function.

Genetic testing will enable practitioners to identify individuals at risk prostate cancer, at, or even before, birth. Presymptomatic diagnosis of these epilepsies will enable prevention of these disorders.

EXAMPLES

The present invention is further detailed in the following Examples, which are offered by way of illustration and are not intended to limit the invention in any manner. Standard techniques well known in the art or the techniques specifically described below are utilized.

EXAMPLE 1

Linkage Analysis

All participants signed informed consent documents. This research project has the approval of the University of Utah School of Medicine Institutional Review Board. Ninety-seven percent of cancer cases have been confirmed through medical records (and/or through the Utah Cancer Registry for prostate cancer cases diagnosed in Utah). Two-point linkage analysis was performed with the package LINKAGE (Lathrop et al., 1984) using the FASTLINK implementation (Cottingham et al., 1993; Schaffer et al., 1994). The statistical analysis for the inheritance of susceptibility to prostate cancer used a model that assumes age-specific incidence rates from the Utah Cancer Registry, and a relative risk of 2.5 for first-degree relatives. Susceptibility to prostate cancer was assumed due to a dominant allele with a population frequency of 0.003. The details of the model are more thoroughly defined in Neuhausen et al. (1999). Marker allele frequencies were estimated from unrelated individuals present in the pedigrees. Linkage in the presence of heterogeneity was assessed by the admixture test (A-test) of Ott (1986), using HOMOG, which postulates two family types, linked and unlinked. Three-point linkage analysis was performed using VITESSE (O'Connell and Weeks, 1995).

EXAMPLE 2

Physical Mapping

BAC DNA was purified and directly sequenced as previously described (Couch et al., 1996). DNA sequences at the SP6 and T7 ends of isolated BAC clones were used to develop STSs that were used for mapping and contig extension. Greater than 95% sequence coverage of the Figure 1 BAC tiling path was obtained by sequencing plasmid sublibraries generated from these clones. The sequence data obtained were assembled into contigs using Acembly, version 4.3 (U. Sauvage, D. Thierry-Mieg and J. Thierry-Mieg; Centre National de la Recherche Scientifique, France). Subsequently, a complete sequence of this interval was released by the MIT genome center.

EXAMPLE 3

Genetic Localization of *HPC2*

A. Early Studies

A set of high risk prostate cancer kindreds has been collected in Utah since 1990 for the purpose of localization of prostate cancer susceptibility loci. In February 1996, linkage analysis of data from a genome scan performed on a subset of the families noted evidence for linkage with markers on chromosome 17p. Subsequent analysis of more markers in this region of chromosome 17p in a larger set of families has led to strong linkage evidence for a susceptibility gene.

TABLE 1

Chromosome 17p Two-point Linkage Evidence

Marker	17p map position	Heterogeneity Lod Score
D17S786	20.0	4.21
Myr 0022	25.5	3.99
Myr 0088	27.0	3.46
D17S947	31.6	2.32
Myr 0084	31.9	3.02
Myr 0079	32.0	0.99
D17S805	43.6	2.25

The study of specific kindreds with strong evidence of linkage to chromosome 17p allows the definition of a most likely region for the susceptibility locus by identifying the smallest inherited piece of chromosome 17p shared by the prostate cancer cases in the kindred. The minimal genetically defined region is based on a telomeric recombinant in kindred 4325 and a centromeric recombinant in kindred 4320. Kindred 4325 was ascertained from a sibship of early onset prostate cancer cases. There are 6 affected brothers in this family, one of whom also has an affected son. Five of the 6 affected brothers, and the affected son, all share the same piece of chromosome 17p from somewhere below marker myr0065 down to and including marker D17S805. Kindred 4320 was also ascertained from a sibship of early onset prostate cancer cases. In this kindred 3 affected brothers and an affected nephew share a piece of

chromosome 17p from D17S786 down to and including myr0084. Together, the kindred 4325 and kindred 4320 recombinants define a minimal region of about 1 megabase (Figure 2A); this localization is well supported by a larger set of recombinants in both directions.

B. Recent Studies

We originally performed a genome-wide search for prostate cancer predisposition loci using a small set of Utah high risk prostate cancer pedigrees and a set of 300 polymorphic markers. The pedigrees were not selected for early age of cancer onset, but were a subset of families ascertained using the Utah Population Database. The first eight pedigrees analyzed gave suggestive evidence of linkage on chromosome 17p near marker D17S520, although significance was not established. We increased the density of markers in the region and expanded the analysis to 33 pedigrees (Table 2A). Analysis of the additional data, using a dominant model integrated with Utah age-specific incidence, yielded the two-point linkage evidence shown in Table 2B. A maximum two-point LOD score of 4.5 was observed at marker D17S1289, theta = 0.07, and a maximum three-point LOD score of 4.3 was observed using the markers D17S1289 and D17S921. Based on these data, we initiated a positional cloning project, focusing on the interval between D17S1289 and D17S921.

Table 2A

Family Resource Used to Detect Linkage to 17p

Number of pedigrees	3.3
Total number of cases	338
Total number of typed cases	188
Mean number of cases/pedigree (range)	10.2 (2-29)
Mean number of typed cases/pedigree (range)	5.7 (1-16)
Mean age of typed cases at diagnosis (range)	68.3 (35-88)

Table 2B

Two-point LOD Scores Using Utah Age-specific Model

Marker	distance (cM)†	Max LOD¥ (theta)	Heterogeneity LOD	(alpha, theta)
D17S796	---	0.11 (.37)	0.10	(1.00, 0.4)
D17S952	10.2	0.90 (.17)	0.87	(1.00, 0.2)
D17S786	10.4	0.00 (.50)	0.95	(0.20, 0.0)
D17S945	12.7	0.38 (.28)	1.41	(0.25, 0.0)
D17S520	15.0	0.69 (.26)	0.64	(1.00, 0.3)
D17S974	15.1	1.01 (.20)	1.20	(0.40, 0.01)
D17S1289	15.2	4.53 (.07)	4.43	(1.00, 0.1)
D17S1159	15.4	0.50 (.27)	1.38	(0.25, 0.0)
GATA134G03	15.7	0.48 (.20)	0.78	(0.75, 0.2)
D17S954	16.2	0.00 (.50)	0.11	(0.40, 0.2)
D17S969	18.2	0.54 (.21)	0.55	(0.85, 0.2)
D17S799	22.0	0.30 (.26)	0.44	(0.70, 0.2)
D17S921	25.2	1.41 (.10)	1.42	(0.95, 0.1)
D17S953	29.2	1.04 (.25)	0.94	(1.00, 0.3)
D17S925	31.2	0.02 (.45)	0.00	(1.00, 0.0)
D17S798	36.2	0.02 (.43)	0.02	(1.00, 0.4)

† Distances estimated from data using CRIMAP (Lander and Green, 1987).

¥ Maximum LOD scores interpolated using the standard quadratic function.

In order to refine the localization of the implied susceptibility gene, we expanded to the set of 127 families (Table 3) which have now been typed at both this locus and the *HPC1* locus. Although the overall data set neither provides significant LOD score evidence for linkage on chromosome 17 nor provides sufficient evidence for *de novo* identification of the *HPC1* locus (Neuhausen et al., 1999), complete haplotyping of the pedigree resource revealed a similar number of prostate cancer-associated haplotypes at each locus.

Table 3

Summary of Resource Genotyped for the Association Tests

Number of pedigrees	127
Total number of cases	2,402
Total number of typed cases	700
Total number of typed pedigree unaffecteds	3,295
Total number of typed divergent controls	243
Mean number of cases/pedigree (range)	18.3 (3-74)
Mean number of typed cases/pedigree (range)	5.5 (1-34)
Mean age of typed cases at diagnosis (range)	66.5 (39-88)

Early in our analysis, we observed that at both 17p and *HPC1* many of our pedigrees segregate haplotypes that are shared by four or more cases, but also contain enough noncarrying cases with respect to either locus to eliminate any linkage evidence within the pedigree, as estimated by LOD score. For instance, 12 affected individuals from kindred 4333 share an HPC1 haplotype and 9 affecteds in kindred 4344 share a 17p haplotype, but neither pedigree shows LOD score evidence for linkage at either locus. While we recognize that this phenomenon may be due simply to lack of linkage, we hypothesized that the underlying cause is actually genetic complexity that is greater than the linkage models can accommodate. We subsequently used multipoint haplotyping software (Thomas et al., 2000) to define segregating haplotypes, and then classified those haplotypes into three groups, depending on strength of evidence: group 1 haplotypes, used for both localization and mutation screening, were defined as haplotypes shared by 4 or more cases and giving a LOD score ≥ 1.0 in the pedigree where they were identified, or haplotypes shared by 6 or more cases irrespective of LOD score; group 2 haplotypes, used for mutation screening only, were defined as haplotypes shared by 4 cases with $0.5 < \text{LOD} < 1.0$ in the pedigree where they were identified, or haplotypes shared by 5 cases with $\text{LOD} < 1.0$; and finally, haplotypes that failed to meet either of the above criteria.

Considering group 1 and 2 haplotypes together, evidence at HPC1 and 17p is quite similar: 43 haplotypes at HPC1 versus 42 at 17p and 258 affected haplotype carriers at HPC1 versus 232 at 17p. Focusing on the group 1 haplotypes, evidence at HPC1 is relatively stronger: 26 group 1 haplotypes at HPC1 versus 18 at 17p and an average of 7.2 affected carriers per group 1 haplotype at HPC1 versus 6.6 at 17p. However, there is one other critical difference between the linkage evidence for the two regions. At *HPC1*, meiotic recombinant mapping

using the group 1 haplotypes has thus far failed to define a consistent region. This is also reflected in the ICPCG *HPC1* study (Xu, 2000); in this work, most of the evidence for linkage comes from a combination of the Utah and Hopkins data sets, but the locations with the best evidence for linkage in each of the individual sets map approximately 15 cM apart. In contrast, recombinant mapping in affected carriers of 17p group 1 haplotypes defined a consistent region (Figure 3). As a result, we were able to focus our contig assembly, transcript map development, and mutation screening efforts on an approximately 1 MB interval centered on D17S947 (Figure 3).

One of the genes mapping near D17S947 shares amino acid sequence similarity with members of the NCBI Cluster of Orthologous Groups (Tatusov et al., 1997) COG1234, typified by the uncharacterized *E. coli* ORF elaC and the uncharacterized *S. cerevisiae* ORF YKR079C. On mutation screening this candidate gene from the genomic DNA of prostate cancer cases carrying 17p group 1 haplotypes, a germline frameshift mutation, 1641 insG, was found in a carrier from kindred 4102. Following detection of this frameshift, the gene, which we shall refer to as *ELAC2* because it is the larger of two human genes that we have found that are homologs of *E. coli* elaC, was subjected to careful sequence and intense genetic analyses.

EXAMPLE 4

Contig Assembly and Genomic Sequencing in the Minimal Genetically Defined HPC2 Region

Contig assembly. Given a genetically defined interval flanked by meiotic recombinants, one needs to generate a contig of genomic clones that spans that interval. Publicly available resources, such as the Whitehead integrated maps of the human genome (e.g., the WICGR Chr 17 map) provide aligned chromosome maps of genetic markers, other sequence tagged sites (STSs), radiation hybrid map data, and CEPH yeast artificial chromosome (YAC) clones.

Oligonucleotide primer pairs for the markers located in the interval were synthesized and used to screen libraries of bacterial artificial chromosomes (BACs) to identify BACs in the region. The initial set of markers used was D17S969, WI-2437, WI-2335, D17S947, and D17S799 (Figure 2A). BACs identified with these markers were end-sequenced. PCR primers designed from those end sequences were used as markers to arrange the initial BACs into contigs. The outermost marker from each contig was used in successive rounds of BAC library screening, eventually enabling the completion of a BAC clone contig that spanned the genetically defined interval. A set of overlapping but non-redundant BAC clones that spanned

this interval (Figure 2A) was then selected for use in subsequent molecular cloning protocols such as genomic sequencing.

Genomic sequencing. Given a tiling path of BAC clones across a defined interval, one useful gene finding strategy is to generate an almost complete genomic sequence of that interval. Two types of random genomic clone sublibraries were prepared from each BAC on the tiling path; these were Sau 3A partial digest libraries with inserts in the 5 to 8 kb size range, and random shear libraries with inserts in the 1.0 to 1.5 kb size range. Plasmid DNA from individual clones from the Sau 3A sublibraries sufficient in number to generate an, on average, 1x redundant sequence of each BAC was prepared using an Autogen robotic plasmid preparation machine (Integrated Separation Systems). Insert DNA from individual clones from the random shear sublibraries sufficient in number to generate an, on average, 5x redundant sequence of each BAC, was prepared by PCR with vector primers directly from aliquots of bacterial cultures of each individual clone. The resulting DNA templates were subjected to DNA sequencing from both ends with M13 forward or reverse fluorescent dye-labeled primers on ABI 377 sequencers.

These sequences were assembled into sequence contigs using the program Acem.bly (Thierry-Mieg et al., 1995; Durbin and Thierry-Mieg, 1991). The genomic sequence contigs were placed in a Genetic Data Environment (GDE) (Smith et al., 1994) local database for subsequent similarity searches. Similarities among genomic DNA sequences and GenBank entries - both DNA and protein - were identified using BLAST (Altschul et al., 1990). The DNA sequences were also characterized with respect to short period repeats, CpG content, and long open reading frames.

EXAMPLE 5

Sequence Assembly of the Human HPC2 Gene

A BLASTn (Altschul et al., 1990) search of genomic sequences from BAC 31k12 against dbEST identified two independent sets of human ESTs that, when parsed across the BAC 31k12 genomic sequences, revealed the presence of two independent multi-exon candidate genes, 04CG09 and the HPC2 gene (Figure 2B). A subset of the EST sequences assigned to HPC2 (Table 4) was assembled to produce a tentative partial cDNA sequence for the gene.

TABLE 4

Human ESTs Used to Assemble a Tentative Partial Human HPC2 cDNA Sequence

EST Accession #	Exon Span
AA679618	1→6
Z17886	4→8
W37591	7→12
AA310236	12→16
R55841	15→19
T34216	18→21
AA634909	20→24
AA504412	23→24
R42795	24→polyA

The individual exons of the human HPC2 gene were identified by parsing that tentative cDNA sequence across the BAC 31k12 genomic sequence (see schematics in Figure 2B). After we had identified the HPC2 gene, the MIT genome sequencing completely sequenced another BAC, 597m12, that also contains all of the exons of HPC2 (GenBank accession # AC005277) The sequence of the human HPC2 gene was corrected both by comparison of the sequences of the individual exons from the tentative cDNA assembly to the corresponding genomic sequences of BACs 31k12 and 597m12, and by mutation screening the gene from a set of human genomic DNAs (see Example 8).

The original tentative human HPC2 cDNA sequence contained neither the start codon nor any of the 5' UTR. These were obtained by biotin capture 5' RACE (Tavtigian et al., 1996). Briefly, a biotinylated reverse primer, CA4cg07.BR2, was designed from the sequence of the third exon of the human HPC2 gene and used, along with the anchor primer 5ampA, for a first round of PCR amplification from human fetal liver cDNA that had been prepared such that the 5' ends of cDNA molecules are anchored with the sequence 5tag1. The resulting PCR products were captured on streptavidin paramagnetic particles (Dynal), washed, and used as template in a second round PCR amplification. A phosphorylated reverse primer, CA4cg07.PR2, was designed from the sequence of the second exon of the human HPC2 sequence and used, along with the nested phosphorylated anchor primer 5ampB, for the second round PCR amplification. The resulting 5' RACE products were gel purified and sequenced with the primer CA4cg07.PR2

using dye-terminator chemistry and ABI 377 sequencers. Analysis of the sequences of these 5' RACE products yielded both the start codon and part of the 5' UTR including an in-frame stop codon (Figure 4). Sequences of the human primers used for 5' RACE are given in Table 5.

A full length human HPC2 cDNA was amplified from human head and neck cDNA using the primers CA4cg7.ATG and CA4cg7.TGA. The cDNA was ligated into the vector pGEM-T Easy (Promega) and transformed into *E. coli*. The sequence of the cDNA clone was confirmed by dye terminator sequencing on ABI 377 sequencers. Sequences of primers used to amplify the cDNA construct and confirm the sequence of the cDNA clone are also given in Table 5.

Table 5

Primers Used in 5' RACE, cDNA Cloning and
Sequence Confirmation of a Full-length Human HPC2 cDNA

5'RACE PRIMERS	Sequence (SEQ ID NO:)
5tag1	CAG GAA TTC AGC ACA TAC TCA TTG TTC Agn n (29)
5AmpA	CAG GAA TTC AGC ACA TAC TCA (30)
5AmpB	(P)TT CAG CAC ATA CTC ATT GTT CA (31)
CA4cg07.BR2	(B)TG AAC GCC TTC TCC ACA GT (32)
CA4cg07.PR2	(P)GT ACC CGC TGC CAC CAC (33)
EXPRESSION CONSTRUCT PRIMERS	
CA4cg7.ATG	GCT AGG ATC CGC CAC CAT GTG GGC GCT TTG CTC (34)
CA4cg7.TGA	GCT ACT CGA GTC ACT GGG CTC TGA CCT TC (35)
SEQUENCING PRIMERS	
M13F20	GTA AAA CGA CGG CCA GT (36)
M13R20	GGA AAC AGC TAT GAC CAT G (37)
CA4cg7F1	TGC GCA CGC GAG AGA AG (38)
CA4cg7R1	CGC TTC TCT CGC GTG CG (39)
CA4cg7F2	TCT AAT GTT GGG GGC TTA (40)
CA4cg7R2	TAA GCC CCC AAC ATT AGA (41)
CA4cg7F3	TGA AAA TGA GCC ACA CCT (42)
CA4cg7R3	AGG TGT GGC TCA TTT TCA (43)
CA4cg7F4	CAT TCA ACC CAT CTG TGA (44)
CA4cg7R4	TCA CAG ATG GGT TGA ATG (45)
CA4cg7F5	TGA ATG CCT CCT CAA GTA (46)
CA4cg7R5	TAC TTG AGG AGG CAT TCA (47)
CA4cg7F6	GCT ACT GGA CTG TGG TGA (48)
CA4cg7R6	TCA CCA CAG TCC AGT AGC (49)
CA4cg7F7	TGG AAG AGT TTC AGA CCT G (50)
CA4cg7R7	CAG GTC TGA AAC TCT TCC A (51)

CA4cg7F8	CGC AGG GAC GCA CCA TA (52)
CA4cg7R8	GGT TGA ACT CGG AGA AGA (53)
CA4cg7F9	CAA CTG GAA AAA TAC CTC G (54)
CA4cg7F10	GCA GAG TCC AGA AAG GC (55)
CA4cg7F11	AGA GGA AAC TTC TTG GTG C (56)
CA4cg7F12	ACC AAG GAA AGG CAG ATG (57)
CA4cg7F13	GTC AAC ATA AGC CCC GAC (58)
CA4cg7F14	GGC TGC TGT GTT TGT GTC (59)
CA4cg7R14	GAA GGC ATT TGG CAG GA (60)
CA4cg7F15	TAT GAT TCC TGC CAA ATG (61)
CA4cg7R15	TCC AGC CAG AGG TGT GC (62)
CA4cg7F16	TGC GAG GCT CTG GTC CG (63)
CA4cg7R16	GGG CAT TGT TGG AAA GTC (64)
CA4cg7F17	TGT TTG CTG GCG ACA TC (65)

nn – the last 2 nucleotides of the anchor sequence 5tag1 are specific for each cDNA prep.
(P) indicates phosphate at the 5' end of the oligo
(B) indicates biotin at the 5' end of the oligo

EXAMPLE 6

Sequence Assembly of the Mouse HPC2 Gene

A BLAST search of the assembled HPC2 cDNA sequence against dbEST identified 5 mouse ESTs that derived from a very similar gene, the mouse ortholog of HPC2, Mm.HPC2; their accession numbers are listed in Table 6.

Table 6

Mouse ESTs Used to Assemble a Tentative Partial Mm.HPC2 cDNA Sequence

EST Accession #	Exon Span
AA563096	1→5
AA518169	8→14
AI132016	16→17
AA184645	19→24
AA174437	24→24

The original partial Mm.HPC2 cDNA sequence contained the start codon but little of the 5' UTR. More extensive 5' UTR sequence was obtained by 5' RACE. Briefly, a biotinylated reverse primer, m04cg07BR1, was designed from the sequence of the fourth exon of the mouse HPC2 gene and used, along with the anchor primer 5ampA, for a first round of PCR amplification from mouse embryo cDNA that had been prepared such that the 5' ends of cDNA

molecules are anchored with the sequence 5tag1. The resulting PCR products were captured on streptavidin paramagnetic particles (Dynal), washed, and used as template in a second round PCR amplification. A phosphorylated reverse primer, m04cg07PR1, was designed from the sequence of the third exon of the mouse HPC2 sequence and used, along with the nested phosphorylated anchor primer 5ampB, for the second round PCR amplification. The resulting 5' RACE products were gel purified and sequenced with the primers m04cg07PR1 and m04cg07 exon2 rev using dye-terminator chemistry and ABI 377 sequencers. Analysis of the sequences of these 5' RACE products yielded both the start codon and part of the 5' UTR including an in-frame stop codon (Figure 4). Sequences of the primers used for 5' RACE are given in Table 7.

More extensive 5' UTR sequence, sequence that may be from the promoter, and the sequences of intron 1 and intron 2 of the mouse HPC2 gene were obtained by genomic sequencing. BAC 428n12 was obtained from a mouse genomic library by screening the library by PCR with a pair of primers (04CG7.m11f1 and 04CG7.m11r1, Table 7) derived from exon 11 of the mouse HPC2 cDNA sequence. A primer pair derived from the SP6 end sequence of BAC 428n12 (428n12.S6.F1 and 428n12.S6.F1, Table 7) was used to screen the mouse BAC library by PCR; several overlapping BACs, including BAC 199n11, were identified. BACs 428n12 and 199n11 were sequenced with a series of 13 sequencing primers (mcg7f1 to mcg7r7, Table 7) derived from mouse HPC2 cDNA dye-terminator chemistry and ABI 377 sequencers. A subset of these sequences were assembled into a genomic sequence contig extending from 280 bp upstream of the ATG start codon of exon 1 into exon 3.

A full length mouse HPC2 cDNA is amplified from mouse embryo, placenta, or fetal brain cDNA using the primers msCA4cg7.f out and msCA4cg7.r out The cDNA is reamplified with the primers msCA4cg7.ATG and msCA4cg7.TGA. The resulting PCR products are gel purified, ligated into the vector pGEM-T Easy (Promega), and transformed into *E. coli*. The sequence of the cDNA clone are confirmed dye terminator sequencing on ABI 377 sequencers. Sequences of primers in use to amplify the cDNA construct are also given in Table 7.

Table 7

Primers Used in 5' RACE and cDNA Cloning of a Full-length Mouse HPC2 cDNA

5'RACE PRIMERS	Sequence (SEQ ID NO:)
5tag1	CAG GAA TTC AGC ACA TAC TCA TTG TTC Agn n (66)
5AmpA	CAG GAA TTC AGC ACA TAC TCA (67)
5 Amp B	(P)TT CAG CAC ATA CTC ATT GTT CA (68)
m04cg07BR1	(B)CA GAA CAC ATT TGG GAA GC (69)
m04cg07PR1	(P)GA TGT TGT CCA AGC GAG C (70)
BAC library screening primers	
04CG7.m11f1	TGA CAC ACA GCA CCT GA (71)
04CG7.m11r1	GAA GAT GTC AGG GTG GA (72)
428n12.S6.F1	CAG GCA TAC CAC TAC AGA (73)
428n12.S6.R1	TAT CAA CTT CTA GGC AAG TG (74)
Genomic sequencing primers	
mcg7f1	GCA CCA TGT CGC AGG GTT C (75)
mcg7r1	GAA CCC TGC GAC ATG GTG C (76)
mcg7f2	TCG CAG GGT TCG GCT CGT C (77)
mcg7r2	AAC CCT GCG ACA TGG TGC G (78)
mcg7f3	AAA GAC CCA CTG CGA CAC C (79)
mcg7r3	GCA GGT GTC GCA GTG GGT C (80)
mcg7f4	CCG AAC ACC GTG TAC CTG CA (81)
mcg7r4	CAG GTA CAC GGT GTT CGG G (82)
mcg7f5	GTC TTC TCG GAA TAC AAC AGG (83)
mcg7r5	CTG TTG TAT TCC GAG AAG AC (84)
mcg7f6	AAG GCG TCC AAC GAC TTA TG (85)
mcg7r6	AGT CGT TGG ACG CCT TCT CC (86)
mcg7r7	TCC GAG TCA GAA AGA TGT TG (87)
EXPRESSION CONSTRUCT PRIMERS	
PRIMARY PCR	
msCA4cg7.f out	GCC TTG TCA GCC TGG TG (88)
msCA4cg7.r out	AGG AAG TGA GCA GAG CG (89)
SECONDARY PCR	
msCA4cg7.ATG	GCT AAA GCT TGC CAC CAT GTG GGC GCT CCG CTC (90)
msCA4cg7.TGA	GCT ACT CGA GTC ACA CTC GCG CTC CTA (91)
SEQUENCING PRIMERS	
m04cg07 exon2 rev	GCC TTC TCC GCA GTT A (92)

nn – the last 2 nucleotides of the anchor sequence 5tag1 are specific for each cDNA prep.
(P) indicates phosphate at the 5' end of the oligo
(B) indicates biotin at the 5' end of the oligo

EXAMPLE 7

Northern Blots

Prehybridization and hybridization were performed at 42°C in 50% formamide, 5x SSPE, 1.0% SDS, 5x Denhardt's mixture, 0.2 mg/mL denatured salmon sperm DNA, and 2 μg/mL poly(A). Dextran sulfate (4% v/v) was included in the hybridization solution only. The membranes were washed twice in 2x SSC/0.1% SDS at 20°C for 30 minutes, followed by a stringency wash in 0.1x SSC/0.1% SDS at 50°C for 30 minutes.

EXAMPLE 8

Mutation Screening of the Human HPC2 Gene

Using genomic DNAs from prostate kindred members, prostate cancer affecteds, and tumor cell lines as templates, nested PCR amplifications were performed to generate PCR products to screen for mutations in the HPC2 gene. The primers listed in Table 8 were used to amplify segments of the HPC2 gene. Using the outer primer pair for each amplicon (1A-1P, i.e., forward A and reverse P of amplicon 1), 10-20 ng of genomic DNA were subjected to a 25 cycle primary amplification, after which the PCR products were diluted 45-fold and reamplified using nested M13-tailed primers (1B-1Q, 1C-1R i.e., nested forward B and nested reverse Q of amplicon 1 or nested forward C and nested reverse R of amplicon 1) for another 23 cycles. In general, samples were amplified with Taq Platinum (Life Technologies) DNA polymerase; cycling parameters included an initial denaturation step at 95°C for 3 min, followed by cycles of denaturation at 96°C (12 s), annealing at 55°C (15 s) and extension at 72°C (30-60 s). After the PCR reactions, excess primers and deoxynucleotide triphosphates were digested with exonuclease I (United States Biochemicals) and shrimp alkaline phosphatase (Amersham). PCR products were sequenced with M13 forward or reverse fluorescent (Big Dye, ABI) dye-labeled primers on ABI 377 sequencers. Chromatograms were analyzed for the presence of polymorphisms or sequence aberrations in either the Macintosh program Sequencher (Gene Codes) or the Java program Mutscreen. We obtained more than 95% double strand sequence coverage for the entire open reading frame of all samples screened.

Table 8

Primers Used to Mutation Screen the HPC2 Gene from Genomic DNA

Exon/Primer name	Sequence (SEQ ID NO:)
HPC2 exon 1	
ca4cg7.m1Anew	CCG CTT GAG ACG CTC TAG TAT (93)
ca4cg7.m1P	GCT CCG AAA GTG CTG ACA G (94)
ca4cg7.m1Bnew	GTT TTC CCA GTC ACG ACG TTT CTA TTG GAT GAG CAG CCT (95)
ca4cg7.m1Qnew	AGG AAA CAG CTA TGA CCA TGC CTG CGA TAT GGT GCG TC (96)
ca4cg7.m1C	GTT TTC CCA GTC ACG ACG CTC AGT TTT GGT GGA GAC G (97)
ca4cg7.m1Rnew	AGG AAA CAG CTA TGA CCA TGT GCC CCG ATG CTC AGA G (98)
HPC2 exons 2&3	(primary)
ca4cg7.m2&23 A2	AAT GGT GTC AGA GAG TTT ACA G (99)
ca4cg7.m2&23P	GCT ATT TGG GAG GCT GAG G (100)
HPC2 exon 2	(nested)
ca4cg7.m2B	GTT TTC CCA GTC ACG ACG AAT GGT GTC AGA GAG TTT ACA G (101)
ca4cg7.m2Q	AGG AAA CAG CTA TGA CCA TGA ACA AGG ACC ACT TTT GCT AT (102)
HPC2 exon 3	(nested)
ca4cg7.m23B	GTT TTC CCA GTC ACG ACG TTT ATA GCA AAA GTG GTC CTT G (103)
ca4cg7.m23Q	AGG AAA CAG CTA TGA CCA TGA GAC TTC CCA CCA GCC TC (104)
HPC2 exon 4	
ca4.cg07.m24A	CCT TGC TGC TTC ACC CTA G (105)
ca4.cg07.m24P	TGC TTT ATA TGT GCT GCT ACG (106)
ca4.cg07.m24B	GTT TTC CCA GTC ACG ACG CAT CTT CCC TGG TTG TAC TTC (107)
ca4.cg07.m24Q	AGG AAA CAG CTA TGA CCA TCT GGA GGG CAG AAG ACT GAT (108)
HPC2 exon 5	
ca4cg7.m3A	CTA CAT TTG TTC AAC CAT AAC TG (109)
ca4cg7.m3P	GAT TTT GAG GTT TGA TGT TGA TG (110)
ca4cg7.m3B	GTT TTC CCA GTC ACG ACG CAT TTG TTC AAC CAT AAC TGC (111)
ca4cg7.m3Q	AGG AAA CAG CTA TGA CCA TAT TTG AGA GGT CAG GGC ATA (112)
HPC2 exon 6	
ca4cg7.m4A	TCG TGT CAG ATT CCC ACC ATA (113)
ca4cg7.m4P	AGG CAT AAG TCA GAC ATC CGT (114)
ca4cg7.m4B	GTT TTC CCA GTC ACG ACG GTT ACT CTT CCC ACA CAT CTT C (115)
ca4cg7.m4Q	AGG AAA CAG CTA TGA CCA TCA CAG CAA GTG TTC AGT TTC TA (116)
HPC2 exon 7	
ca4cg7.m5A	CAT TCC CAT GTA TGA ACG TCT (117)
ca4cg7.m5P	ATA GTA AGC CCA GGA AGA AGGA (118)

ca4cg7.m5B	GTT TTC CCA GTC ACG ACG CAT TCC CAT GTA TGA ACG TCT (119)
ca4cg7.m5Q	AGG AAA CAG CTA TGA CCA TCT ACA AGC ATT ACA AGG CAG AG (120)
HPC2 exon 8	
ca4cg7.m6A	AGT GTC TTC AGC CTT TGT ATT G (121)
ca4cg7.m6P	ATC TGC TAT CTC TTC TTG TCT CA (122)
ca4cg7.m6B	GTT TTC CCA GTC ACG ACG ATC GGG TCA TAA TCA GTC TGT G (123)
ca4cg7.m6Q	AGG AAA CAG CTA TGA CCA TAT CTC TTC TTG TCT CAG GTA ACA (124)
HPC2 exons 9&10	(primary)
ca4cg7.m7&8A	CTT CTG AAA GCA ATA AAC GCA T (125)
ca4cg7.m7&8P	GAT GTC CAA ACT GTT CCA CG (126)
HPC2 exon 9	(nested)
ca4cg7.m7B	GTT TTC CCA GTC ACG ACG TAA AAC CAA CCT TCT TCA TTA G (127)
ca4cg7.m7Q	AGG AAA CAG CTA TGA CCA TAG CAA TGA TGG GAG CGA TG (128)
HPC2 exon 10	(nested)
ca4cg7.m8B	GTT TTC CCA GTC ACG ACG GGC TTC TGG GGA CTC ACT G (129)
ca4cg7.m8Q	AGG AAA CAG CTA TGA CCA TCC TTC AAA AGT GGT GTC TGT AG (130)
HPC2 exon 11	
ca4.cg07.m9A	GTA TCC ACA AAG AGA CCA GAA G (131)
ca4.cg07.m9P	CAC CAA CTA CCA ACA GTG ACT TA (132)
ca4.cg07.m9B	GTT TTC CCA GTC ACG ACG GCT CAC TGG ATA GGA TAT GTC AT (133)
ca4.cg07.m9Q	AGG AAA CAG CTA TGA CCA TCC AGA AAC ACA GCT CTT GCC (134)
HPC2 exon 12	
ca4.cg07.m10A	GCT TGC CAG ATA CAG GAA TC (135)
ca4.cg07.m10P	ACA GAA AGT TTA GGC AGG TG (136)
ca4.cg07.m10B	GTT TTC CCA GTC ACG ACG ACG ATA CCC CTC CCT GGC T (137)
ca4.cg07.m10Q	AGG AAA CAG CTA TGA CCA TAC AGA AAG TTT AGG CAG GTG (138)
HPC2 exons 13&14	(primary)
ca4.cg07.m11&12A	CCT CTC ACT CTT CCC AGC AC (139)
ca4.cg07.m11&12P	GGA GTA GGC TGC TTT TCT AAA T (140)
HPC2 exon 13	(nested)
ca4.cg07.m11B	GTT TTC CCA GTC ACG ACG GAA CAC CTC ATC CTC ATT ACC A (141)
ca4.cg07.m11Q	AGG AAA CAG CTA TGA CCA TAA GAG ACA AAA CAC ATT CAT GG (142)
HPC2 exon 14	(nested)
ca4.cg07.m12B	GTT TTC CCA GTC ACG ACG GTT TCC GCT GTA AGG TAG TGT (143)
ca4.cg07.m12Q	AGG AAA CAG CTA TGA CCA TCT GGA ACA TTT ACT ATG TGG CTA (144)

HPC2 exon 15	
ca4.cg07.m13A	TGC TAG TGG GTA GAG GTC AG (145)
ca4.cg07.m13P	ACT GAA AGC CAG GTT AGA ATG (146)
ca4.cg07.m13B	GTT TTC CCA GTC ACG ACG ACC CTG TCC GTC ACC TGA G (147)
ca4.cg07.m13Q	AGG AAA CAG CTA TGA CCA TCC CAC CAG CAC TCC ACT TA (148)
HPC2 exon 16	
ca4cg07.m14A	TGT GAA GAC GGG ATA ACC TGA (149)
ca4cg07.m14P	GAC AGG GCT TGA TAC CGCA (150)
ca4cg07.m14B	GTT TTC CCA GTC ACG ACG ATG CTG GCT CAC TTT TGA CC (151)
ca4cg07.m14Q	AGG AAA CAG CTA TGA CCA TGAC TGG TGA GTA CAG CAG GA (152)
HPC2 exon 17	
ca4.cg07.m15A	CCA GCC TTT GTG TAA GTC TAC (153)
ca4.cg07.m15P	TCT GGG CAA GTT TGG AAG C (154)
ca4.cg07.m15B	GTT TTC CCA GTC ACG ACG TCC AAA GCA GAC ATC AGC CTC (155)
ca4.cg07.m15Q	AGG AAA CAG CTA TGA CCA TGG AGG AAA AGA CGC AGC CA (156)
HPC2 exon 18	
ca4.cg07.m16A	CGC TTT CTG CCT GTG ACA T (157)
ca4.cg07.m16P	TTC TGT CCT TCA GCC AAT GC (158)
ca4.cg07.m16B	GTT TTC CCA GTC ACG ACG TTA GAG GCT GGT GGG TGA C (159)
ca4.cg07.m16Q	AGG AAA CAG CTA TGA CCA TCA TCT CAA TAA AAA CTG GAG TGC (160)
HPC2 exon 19	
ca4.cg07.m17A	CAC TTG ATG GGC GTT CTG AG (161)
ca4.cg07.m17P	TTC TGT CCT TCA GCC AAT GC (162)
ca4.cg07.m17B	GTT TTC CCA GTC ACG ACG TTC CAG CGG TTT ACA CAT CA (163)
ca4.cg07.m17Q	AGG AAA CAG CTA TGA CCA TTA CCC CAG TGT CCA CCT TG (164)
HPC2 exons 20&21	(primary)
CA4CG7.m18&22A	GGG TTC TCC AGC CAA AGA CT (165)
CA4CG7.m18&22P	CTG AGT CTC CTG CCT CTG C (166)
HPC2 exon 20	(nested)
ca4.cg07.m18B	GTT TTC CCA GTC ACG ACG GGG TTC TCC AGC CAA AGA CT (167)
ca4.cg07.m18Q	AGG AAA CAG CTA TGA CCA TGT GGG GCT GGA AGG CTC TG (168)
HPC2 exon 21	(nested)
ca4.cg07.m22B	GTT TTC CCA GTC ACG ACG AAG AGG TAA GGG GCA CAG C (169)
ca4.cg07.m22Q	AGG AAA CAG CTA TGA CCA TCT GAG TCT CCT GCC TCT GC (170)
HPC2 exon 22	
ca4.cg07.m19A	GCT GAG TGT TGA GAC CAG GA (171)
ca4.cg07.m19P	AGA CAA ACG ACG GCT GCT C (172)

ca4.cg07.m19B	GTT TTC CCA GTC ACG ACG TTG AGA CCA GGA AAC AGC AC (173)
ca4.cg07.m19Q	AGG AAA CAG CTA TGA CCA TGA GAG GAT GTG GGC GAC AA (174)
HPC2 exon 23	
ca4.cg07.m20A	GGG AGA TGG TGC TGG CTA C (175)
ca4.cg07.m20P	CCT GGT TAG TGA TGG GTA GAT (176)
ca4.cg07.m20B	GTT TTC CCA GTC ACG ACG CAG GGT CTG TGC CAC TGT C (177)
ca4.cg07.m20Q	AGG AAA CAG CTA TGA CCA TCT CAG TGT GTA GAG TCC TGT C (178)
HPC2 exon 24	splice acceptor and open reading frame
ca4.cg07.m21A	TTG ATT TTG AGA GCA TCT GGA C (179)
ca4.cg07.m21P	CTC GGA CAC TTA GAC CCA CTG (180)
ca4.cg07.m21B1	GTT TTC CCA GTC ACG ACG TGC ATC CCT TCC AGC TCC T (181)
ca4.cg07.m21Q	AGG AAA CAG CTA TGA CCA TGA CAC ACA GCC TTC TGA GTT CA (182)
ca4.cg07.m21C	GTT TTC CCA GTC ACG ACG CCA CAC AGA GGA GCC ACA G (183)
ca4.cg07.m21R	AGG AAA CAG CTA TGA CCA TAC CAG TCC TAA GAG GCA TCT ATA (184)
HPC2 exon 24	3' untranslated region
ca4.cg07.m21.3'UTR A	CCA CAC AGA GGA GCC ACA G (185)
ca4.cg07.m21.3'UTR P	CCA GAG GTG CTC ACT ACG AC (186)
ca4.cg07.m21.3'UTR B	GTT TTC CCA GTC ACG ACG AGG TCA GAG CCC AGT GAA GAT (187)
ca4.cg07.m21.3'UTR Q	AGG AAA CAG CTA TGA CCA TCA TCT GCT TGC TTC CGT GTG (188)
ca4.cg07.m21.3'UTR C	GTT TTC CCA GTC ACG ACG TCA GGA TAG GTG GTA TGG AGC (189)
ca4.cg07.m21.3'UTR R	AGG AAA CAG CTA TGA CCA TCG GAC ACT TAG ACC CAC TGA T (190)

Table 9

Sequence Variants

Variant name	Sequence (SEQ ID NO:)	Coding effect*
C650T	AGACTCCGAGTYGAATGAAAATG (191)	Ser217Leu
A1560G	GGTGAGGGCACRTTTGGGCAGCT (192)	Thr520Thr
G1621A	GCACCCTGGCTRCTGTGTTTGTG (193)	Ala541Thr
1641insG (normal)	GTGTCCCACCTG-CACGCAGATCA (194)	
(with insertion of G)	GTGTCCCACCTGGCACGCAGATCA (195)	frameshift
C1722T	AAGCCGCTTCAYCCTTTGCTGGT (196)	His574His
A1893G	GCTGTTGCGAACRTGTGATTTGGA (197)	Thr631Thr
C2632G	GAGGCTTGGGSTCCCACATAAG (198)	
C2687T	CCTGGCACAGCYGCGGGCCAGGA (199)	
G2801A	AATCCAGCAAARTGATTCCCTGC (200)	
IVS2 T-11C	Taaatgttttytcattcttag (201)	
IVS5 T-14C	Ttgctgttgtgyggttttcttgt (202)	
IVS10 23insGAT (normal)	ggttttcttgat---tcagcagttaca (203)	
(with insertion of GAT)	Ggttttcttgatgattcagcagttaca (204)	
IVS13 C15T	Gtgtctcagacyggccccttgtc (205)	
IVS14 A17T	Tgccatcttgawctaatggaatc (206)	
IVS14 T-8C	Cttctctctcctycctgcagggat (207)	
IVS16 C41T	Catcaagggcaygtttacttttt (208)	
IVS19 C26G	Cagccttgcccsctgggctgttg (209)	

*based on conceptual translation of the HPC2 ORF for each allele of the sequence variant.

Kindred 4102 was ascertained as a high risk cluster with eight prostate cancer cases in a three generation pedigree. Genotyping revealed that six of the eight cases shared a chromosome 17p haplotype. The youngest (age at diagnosis of 46) affected carrier of this shared haplotype, 4102.013 (i.e., kindred #4102, individual #013; Figure 5A), was selected for mutation screening. On mutations screening lymphocyte DNA from 4102.013, we detected a frameshift, 1641 insG, in *ELAC2*. A test for segregation revealed that the frameshift was not on the father's chromosome, but rather was inherited through the carrier's mother, 4102.002. Her affected uncle 4102.053 was diagnosed with and died of prostate cancer at age 76 in the 1960s. Genotyping of his children demonstrated that he was an obligate frameshift carrier. In all, there are five male frameshift carriers over age 45 in the pedigree. Of these, three have prostate cancer, the fourth has a PSA of 5.7 at age 71, and the fifth has a PSA of 4.2 at age 74 (Figure

5A). The frameshift occurs at His 548, within the histidine motif (Figures 6A-B) and is predicted to be quite disruptive to the protein.

As the frameshift 1641 insG was found in an individual with early onset prostate cancer, we screened an additional 45 prostate cancer cases with early age at diagnosis (Dx ≤ 55 years), irrespective of evidence of linkage to any locus, for mutations in *ELAC2*. An alteration, Arg 781 His, was identified in individual 4289.003, diagnosed with prostate cancer at age 50. Upon expansion of his pedigree, the mutation was traced back four generations to 4289.006, who had affected descendants from five known wives. Prostate cancer cases who carry the missense change have been found among the descendants from three of these five marriages. Of thirteen prostate cancer cases in the pedigree, six carry the missense change, three are unknown, and four are non-carriers. In addition, a female carrier of this missense change, 4289.183, was diagnosed with ovarian cancer at age 43 (Figure 5B). Within the generations with phenotype information, there are only two unaffected male mutation carriers over age 45; 4289.068 (PSA of 0.6 at age 60) and 4289.063, who died of a heart attack at age 62. We have no additional information on 4289.063; however, two of 4289.063's sons and a grandson are carriers who have been diagnosed with prostate cancer. The missense change occurs in a very highly charged stretch of amino acid residues near the C-terminus of the protein. Arg 781 is conserved in mouse (Figures 6A-B), and the charge character of the sequence segment is conserved in *C. elegans*. While one cannot definitively predict that this missense change will affect protein function, expansion from a single affected mutation carrier to a pedigree with a LOD score of 1.3 provides good evidence that the mutation is in fact deleterious.

The identification of two mutations provides strong evidence that *ELAC2* is a prostate cancer susceptibility gene. However, after screening 42 haplotypes with evidence for linkage at 17p, we have found only these two high-risk mutations. Thus it seems that only a small fraction of prostate cancer pedigrees segregate obvious mutations in the *ELAC2* coding sequence. We do not yet know what fraction of the pedigrees harbor subtle gene rearrangements or regulatory mutations.

Taken together, the observation that the frameshift HPC2 1641insG segregates with prostate cancer across three generations of kindred 4102, and the inference from shared sequence similarity that the frameshift HPC2 1641insG must be deleterious to the function of the HPC2 protein, establish that deleterious germline mutations in the HPC2 gene confer susceptibility to prostate cancer.

EXAMPLE 9

Common Missense Changes in HPC2

When our original set of linked pedigrees was screened for mutations in *ELAC2*, we observed several occurrences of the non-conservative missense change Ser 217 Leu. This missense change is embedded in an extremely hydrophilic segment of the protein sequence. Like the common human allele, the mouse and *C. elegans* residues at this position are also serine. Although the sequence of this segment is not well conserved, its hydrophilic character is (Figures 6A-B); thus substitution of a bulky hydrophobic residue for Ser 217 could result in structural consequences to the protein.

We analyzed this sequence variant in our pedigree cases, unaffected pedigree members, and an unrelated set of males who have no diagnosis of cancer (divergent controls). The total number of individuals typed exceeded 4,000 (Table 3), with an overall allele frequency of 30% for Leu 217. A logistic regression was performed for disease status to delineate effects of genotypes at Ser 217 Leu versus birth year (a demographic datum collected on all participants). We observed a significant interaction between genotype and birth year ($p = 0.027$), indicating that association tests should be performed which appropriately considered birth cohort. Figure 7 illustrates this birth effect, showing that genotype frequencies differ across birth cohorts for cases, but appear more uniform for the unaffected controls. We subsequently chose to analyze the effect of genotype in individuals born after 1919, since the data suggest that a different risk pattern may exist for individuals born before this date.

Association tests are consistent with the hypothesis that the Leu 217 variant is deleterious or in disequilibrium with another deleterious variant. Prostate cancer patients born between 1920 and 1959 have a significantly higher proportion of Leu 217 homozygotes than either the divergent controls (57/429 vs. 9/148, p-value = 0.026) or the unaffected pedigree members (57/429 vs. 220/2371, p-value = 0.013) (Figures 7 and 8). That Leu 217 is so common could be explained by the allele contributing to a common disease in a recessive manner.

Upon mutation screening *ELAC2* in the set of early onset prostate cancer cases, we also observed several occurrences of a second non-conservative missense change, Ala 541 Thr. This missense change occurs at the border of the histidine motif (Figures 6A-B and 9) and thus may well affect the protein's function. This variant has been examined in the same set of cases and controls, where it has an overall allele frequency of 4%. Thr 541 is in strong disequilibrium with Leu 217; in fact, we have yet to observe a chromosome that carries Thr 541 that does not also carry Leu 217. Another logistic regression was performed to investigate effects of

genotypes at Ala 541 Thr. Again, a significant interaction between genotype and birth year was found (p = 0.003), along with evidence for an effect of genotype at Ala 541 Thr on disease status. Table 10 shows the allele frequencies.

Table 10

Allele Definitions

Allele	Defining Sequence Variant(s)	Note
0	wt	Matches mouse at polymorphic positions
1	Leu 217	Allele frequency = 26.0%
2	Leu 217 + Thr 541	Allele frequency = 3.9%

The carrier frequency of Thr 541 is significantly higher in prostate cancer cases than divergent controls such that the variant appears to be dominant and deleterious (carrier frequency of 42/429 vs. 5/148, p-value = 0.022) (Figure 8). In contrast, the Thr 541 carrier frequency is not significantly higher in the cases than the unaffected pedigree members. However, in the comparison between cases and pedigree unaffecteds, when Leu 217 homozygotes are subdivided into Thr 541 carriers and non-carriers, the presence of Thr 541 is associated with a higher odds ratio (2.0 vs. 1.4) and the model remains statistically significant (p-value = 0.017, trend test p-value 0.004) (Figure 8). Thus both comparisons support the hypothesis that the allele bearing both Thr 541 and Leu 217 is more deleterious than the allele bearing just Leu 217.

EXAMPLE 10

Identification of HPC2-interacting Proteins by Two-hybrid Analysis

DNA fragments encoding all or portions of HPC2 are ligated to a two-hybrid DNA-binding domain vector such as pGBT.C such that the coding sequence of *HPC2* is in-frame with coding sequence for the Gal4p DNA-binding domain. A plasmid that encodes a DNA-binding domain fusion to a fragment of HPC2 is introduced into the yeast reporter strain (such as J692) along with a library of cDNAs fused to an activation domain. Transformants are spread onto 20 - 150 mm plates of selective media, such as yeast minimal media lacking leucine, tryptophan, and histidine, and containing 25 mM 3-amino-1,2,4-triazole. After one week incubation at 30° C, yeast colonies are assayed for expression of the *lacZ* reporter gene by β-galactosidase filter assay. Colonies that both

grow in the absence of histidine and are positive for production of β-galactosidase are chosen for further characterization.

The activation domain plasmid is purified from positive colonies by the smash-and-grab technique. These plasmids are introduced into *E. coli* (e.g., DH10B (Gibco BRL) by electroporation and purified from *E. coli* by the alkaline lysis method. To test for the specificity of the interaction, specific activation domain plasmids are cotransformed into strain J692 with plasmids encoding various DNA-binding domain fusion proteins, including fusions to segments of HPC2 and human lamin C. Transformants from these experiments are assayed for expression of the *HIS3* and *lacZ* reporter genes. Positives that express reporter genes with *Hs.HPC2* constructs and not with lamin C constructs encode bona fide HPC2-interacting proteins. These proteins are identified and characterized by sequence analysis of the insert of the appropriate activation domain plasmid.

This procedure is repeated with mutant forms of the HPC2 gene, to identify proteins that interact with only the mutant protein or to determine whether a mutant form of the HPC2 protein can or cannot interact with a protein known to interact with wild-type HPC2.

EXAMPLE 11

Identification and Sequencing of Orthologs and a Paralog of the Human HPC2 Gene

All species living on the Earth now are thought to have evolved from a single common ancestor that lived in the distant past, perhaps 3.5 to 4 billion years ago. This means that any pair of species must share a common ancestor species that lived at some time in the past. Admittedly, this view is a bit simplistic because, for instance, the nuclear genomes and mitochondrial genomes of eukaryotes are thought to have independent prokaryotic ancestries. During the evolution of an ancestral species into two or more extant daughter species, the genes present in the genome of the ancestral species evolve into the genes present in the genomes of the daughter species. The evolutionary history of the genes present in the daughter species can be quite complex because the individual genes can evolve through a diverse set of processes including nucleotide substitution, insertion, deletion, gene duplication, gene conversion, lateral transfer, etc. Even so, the evolutionary history of related genes in related organisms can often be sorted out, especially if the pair/set of species share a relatively recent common ancestor or if the genes being analyzed evolved primarily through nucleotide substitutions and/or small insertions and/or small deletions, but not gene duplications or gene conversions. When, upon analysis, it

appears that a single gene in one species and a single gene in another species have evolved from a single gene in a common ancestor species, those genes are termed orthologs.

Knowledge of the identity of genes orthologous to disease-related human genes can often be quite useful.

The human *HPC2* cDNA sequence was assembled from a combination of ESTs, hybrid selected clones, and 5' RACE (Rapid Amplification of cDNA Ends) products; the orthologous mouse *Elac2* cDNA sequence was assembled from ESTs and 5' RACE products. Conceptual translation of the human cDNA sequence yielded a protein of 826 amino acids; parsing the cDNA sequence across the corresponding genomic sequence revealed 24 coding exons (Figure 3). Mouse *Elac2* encodes a protein of 831 residues in 25 exons. BLAST (Altschul et al., 1990) searches of the *ELAC2* sequence against GenBank readily revealed a single ortholog in *S. cerevisiae* (YKR079C) and a single ortholog in *C. elegans* (CE16965, CELE04A4.4), but two related sequences in *S. pombe* and *A. thaliana*. Alignment of representative family members revealed a block of good conservation near the N-termini and a series of blocks of high similarity across the C-terminal half of the proteins (Figures 6A-B and 10).

Hybridization of RNA blots to labeled fragments of human *ELAC2* cDNA revealed a single transcript of approximately 3 kb (Figures 11A-D), in agreement with our full-length cDNA assembly of 2,970 bp. The transcript was detected in all tissues surveyed and, like *BRCA1* and *BRCA2*, was most abundant in testis. The apparent size of the transcript agrees well with our full length cDNA assembly, 2970 bp. There was no evidence from RNA blots, EST sequences, or RT-PCR experiments of significant alternative splicing of the transcript.

In the course of surveying ESTs derived from this gene, we identified a small number of human and rabbit ESTs derived from a second, related gene. The human cDNA sequence of this related gene was assembled from a combination of ESTs and 5' RACE products. Conceptual translation revealed that the transcript encodes a protein of 363 residues. Radiation hybrid mapping placed the gene at approximately 365 cR on chromosome 18. When this sequence, along with representative sequences from a eubacterium (*E. coli* elaC), a cyanobacterium (*Synechocystis* sp. gi2500943/SLR0050) and an archaebacterium (*M. thermoautotrophicum* gi2622965) was added into the multiprotein alignment (Figures 6A-B), it became apparent that two distinct groups of proteins were represented; a group of larger proteins (800-900 aa) restricted to the eukaryotes, and a group of smaller proteins (300 to 400 aa) that align with the C-terminal half of the former group and includes sequences from the eukaryotes, eubacteria, and

archaebacteria. As the 363 residue human protein falls into this second group and is more similar to *E. coli* elaC than is ELAC2, we will refer to it as *ELAC1*.

The alignment revealed a striking histidine containing motif, ϕϕϕ[S/T]HxHxDHxxG (SEQ ID NO:214), where ϕ can be any large hydrophobic residue, near the N-terminus of the *ELAC1* group, and in the C-terminal portion of the ELAC2 group. This motif is reminiscent of the histidine motif found in the metallo-β-lactamases (Melino et al., 1998) and suggests, in accord with the annotation for COG1234 (www.ncbi.nlm.nih.gov/COG/index.html), that the proteins are metal-dependent hydrolases. While assembling the multiple alignment, we observed that the sequence within which the histidine motif is embedded also aligns with the *ELAC2* N-terminal conserved block (Figure 12), leading us to predict that some structural feature of the protein is repeated. Even so, the N-terminal copy of the repeated sequence would not necessarily retain metal-dependent hydrolase activity, as the histidine motif itself is not conserved.

Thorough BLAST searches of GenBank using sequences containing this histidine motif, combined with iterative motif searches (Nevill-Manning et al., 1998) using the eMOTIF SCAN website (http://dna.stanford.edu/scan), revealed two other families of proteins that share extended amino acid sequence similarity with members of COG1234. The similarity includes 4 to 6 shared motifs distributed across the *ELAC1* domain (Figure 9). One such family is the PSO2 (or SNM1) family of DNA inter-strand crosslink repair proteins (Haase et al., 1989; Meniel et al., 1995; Niegemann and Brendel, 1994), present only in eukaryotes. The second family encodes the 73 kDa subunit of the mRNA cleavage and polyadenylation specificity factor (CPSF73) (Chanfreau et al., 1996; Jenny et al., 1994; Jenny et al., 1996). Surprisingly, members of this latter gene family are present in both eukaryotes and archaebacteria, as well as a cyanobacterium. These three gene families, *ELAC1/2*, PSO2 and CPSF73, are equally similar to each other (Figures 9 and 13); indeed they were originally placed in a single COG (Tatusov et al., 1997). While PSO2 is required for repair of DNA inter-strand crosslinks following treatment of cells with, for instance, 8-methoxypsoralen plus UV-irradiation (Menial et al., 1995), the actual substrate for the protein's presumptive metal-dependent hydrolase activity has not been defined. Similarly, although CPSF73 is a component of the mRNA 3' end cleavage and polyadenylation specificity factor, it has neither the 3' end cleavage nor the polyadenylation activity, and the substrate for its presumptive metal-dependent hydrolase activity is unknown. While the *S. cerevisiae* CPSF73 ortholog YSH1 (BRR5) is an essential gene, PSO2 is not. Given the phylogenetic conservation of the *ELAC1* domain and the observation that *S.*

cerevisiae encodes only a single member of this gene family, YKR079C, we asked whether it is an essential gene. To answer this question, we performed one-step gene disruption of YKR079C using URA3 as a selectable marker in yeast diploid cells. Two heterozygote knockout strains were sporulated and tetrads were dissected. Each tetrad yielded 1 or 2 viable haploid colonies; these were all URA$^-$ and YKR079C wt. Thus we concluded that, like YSH1, YKR079C is an essential gene.

In addition to the histidine motif and the local sequence context in which it is embedded, ELAC1/2, PSO2 and CPSF73 proteins share a series of sequence features, some shared pairwise between the gene families and others by all three. Strikingly, all three families have three or four conserved histidine or cysteine positions, past the histidine motif, that lie within these shared regions and can be aligned across the gene families (Figure 9). The arrangement is reminiscent of the binuclear zinc binding active site of some metallo-β lactamases (Carfi et al., 1998; Fabiane et al., 1998) and the shared similarity between the metallo-β lactamases and glyoxalase II (Melino et al., 1998). This series of sequence similarities leads to three predictions. First, the extended similarity between the ELAC1/2, PSO2 and CPSF73 protein families suggests that they share a domain of approximately 300 residues, and this domain constitutes a metal-dependent hydrolase that coordinates two-divalent cations in its active site. Second, the overall fold of this domain is likely to be similar to that of the metallo-β lactamases. Third, similarity between the region surrounding the ELAC1/2 histidine motif and the N-terminus of the ELAC2 proteins suggests that these proteins are comprised of two structurally similar domains and arose from a direct repeat/duplication of an ancestral *ELAC1*-type gene.

A number of members of the *ELAC1/2* family are annotated in GenBank as sulfatases or sulfatase homologs. The annotation appears to be assigned through sequence similarity to the atsA gene of *Alteromonas carrageenovora*. The atsA protein contains a histidine motif and has been demonstrated to have aryl sulfatase activity *in vitro* (Barbeyron et al., 1995), though its sequence does not contain any of the typical sulfatase motifs listed by PROSITE. No other experimentally verified aryl sulfatase contains the histidine motif. As the *E. coli* protein most similar to *A. carrageenovora* atsA is elaC, atsA may well be a diverged member of the *ELAC1* gene family (BLASTp and alignment not shown). Accordingly, *ELAC1* family members should be tested for aryl sulfatase activity; however, it is not apparent whether *ELAC1* and *ELAC2* family members have the same substrate.

In addition to the paralog and the mouse ortholog mmELAC2 (for *Mus musculus* ELAC2), orthologs of HPC2 have been identified in chimpanzee and gorilla. These are ptELAC2 (*Pan troglodytes* ELAC2) and ggELAC2 (*Gorilla gorilla* ELAC2).

EXAMPLE 12

Multiple Protein Sequence Alignments

For the alignment of Figures 6A-B, shading criteria were identity (white on black) or conservative substitution (white on gray) for all ELAC2 sequences with a residue at that position, with four of the five sequences actually having to have a residue at that position. Shaded positions in the ELAC2 sequences were propagated into the ELAC1 sequences. For the alignment of Figure 12, two shading criteria were used: (1) Identity or conservative substitution across the ELAC2 N-terminal alignment and identity or conservative substitution across either the ELAC1 or ELAC2 His motif. (2) Identity or conservative substitution across both the ELAC1 and ELAC2 His motif, with some conservation across the ELAC2 N-terminal alignment. For the alignment of Figure 9, shading criteria were identity or conservative substitution across two out of the three (CPSF73, PSO2, ELAC2) protein families represented.

EXAMPLE 13

Analysis of the HPC2 Gene

The structure and function of HPC2 gene are determined according to the following methods.

Biological Studies. Mammalian expression vectors containing HPC2 cDNA are constructed and transfected into appropriate prostate carcinoma cells with lesions in the gene. Wild-type HPC2 cDNA as well as altered HPC2 cDNA are utilized. The altered HPC2 cDNA can be obtained from altered HPC2 alleles or produced as described below. Phenotypic reversion in cultures (e.g., cell morphology, doubling time, anchorage-independent growth) and in animals (e.g., tumorigenicity) is examined. The studies will employ both wild-type and mutant forms of the gene.

Molecular Genetics Studies. *In vitro* mutagenesis is performed to construct deletion mutants and missense mutants (by single base-pair substitutions in individual codons and alanine scanning mutagenesis). The mutants are used in biological, biochemical and biophysical studies.

Mechanism Studies. The ability of HPC2 protein to bind to known and unknown DNA sequences is examined. Its ability to transactivate promoters is analyzed by transient reporter expression systems in mammalian cells. Conventional procedures such as particle-capture and yeast two-hybrid system are used to discover and identify any functional partners. The nature and functions of the partners are characterized. These partners in turn are targets for drug discovery.

Structural Studies. Recombinant proteins are produced in *E. coli*, yeast, insect and/or mammalian cells and are used in crystallographic and NMR studies. Molecular modeling of the proteins is also employed. These studies facilitate structure-driven drug design.

EXAMPLE 14

S. cerevisiae Gene Knockout

The URA3 gene was PCR amplified with tailed primers resulting in a product flanked by 42 bp of YKR079C coding sequences (amino acids 3-16 and 818-831). The resulting PCR product was transformed into yeast diploid strain YPH501 (Stratagene); URA^+ clones were screened for disruption by the presence of a shorter PCR product at the YKR079C locus. The knock-out clones were further confirmed by sequencing the shorter PCR product for the presence of URA3 sequences. Two heterozygote knockout strains were sporulated and tetrads dissected. Each tetrad yielded 1 or 2 viable colonies. These were genotyped at YKR079C and tested for growth on URA^- plates.

EXAMPLE 15

Association Tests

STSs for Ser 217 Leu and Ala 541 Thr were amplified by allele specific PCR using fluorescently labeled oligos. Allele calls were made with our automated genotyping system. Genotype calls required good allele calls at both markers. Logistic regression analyses were performed using the SPSS statistical software package. The chi-squared statistics for the 2x2 contingency tables were calculated with the Yates correction. The trend statistic for the 3x2 contingency table was calculated with the Cochran-Armitage trend test (Cochran, 1954; Armitage, 1955) using a simple linear trend (0,1,2) for the row scores.

EXAMPLE 16

Generation of Polyclonal Antibody against HPC2

Segments of HPC2 coding sequence are expressed as fusion protein in *E. coli*. The overexpressed proteins are purified by gel elution and used to immunize rabbits and mice using a procedure similar to the one described by Harlow and Lane, 1988. This procedure has been shown to generate Abs against various other proteins (for example, see Kraemer, *et al.*, 1993).

Briefly, a stretch of HPC2 coding sequence was cloned as a fusion protein in plasmid PET5A (Novagen, Inc., Madison, WI). The HPC2 incorporated sequences might include SEQ ID NOs:1, 3 or 28 or portions thereof. After induction with IPTG, the overexpression of a fusion protein with the expected molecular weight is verified by SDS/PAGE. Fusion proteins are purified from the gel by electroelution. The identification of the protein as the HPC2 fusion product is verified by protein sequencing at the N-terminus. Next, the purified protein is used as immunogen in rabbits. Rabbits are immunized with 100 μg of the protein in complete Freund's adjuvant and boosted twice in 3 week intervals, first with 100 μg of immunogen in incomplete Freund's adjuvant followed by 100 μg of immunogen in PBS. Antibody containing serum is collected two weeks thereafter.

This procedure can be repeated to generate antibodies against mutant forms of the HPC2 protein. These antibodies, in conjunction with antibodies to wild type HPC2, are used to detect the presence and the relative level of the mutant forms in various tissues and biological fluids.

EXAMPLE 17

Generation of Monoclonal Antibodies Specific for HPC2

Monoclonal antibodies are generated according to the following protocol. Mice are immunized with immunogen comprising intact HPC2 or HPC2 peptides (wild type or mutant) conjugated to keyhole limpet hemocyanin using glutaraldehyde or EDC as is well known.

The immunogen is mixed with an adjuvant. Each mouse receives four injections of 10 to 100 μg of immunogen and after the fourth injection blood samples are taken from the mice to determine if the serum contains antibody to the immunogen. Serum titer is determined by ELISA or RIA. Mice with sera indicating the presence of antibody to the immunogen are selected for hybridoma production.

Spleens are removed from immune mice and a single cell suspension is prepared (see Harlow and Lane, 1988). Cell fusions are performed essentially as described by Kohler and Milstein, 1975. Briefly, P3.65.3 myeloma cells (American Type Culture Collection, Rockville,

MD) are fused with immune spleen cells using polyethylene glycol as described by Harlow and Lane, 1988. Cells are plated at a density of $2x10^5$ cells/well in 96 well tissue culture plates. Individual wells are examined for growth and the supernatants of wells with growth are tested for the presence of HPC2 specific antibodies by ELISA or RIA using wild type or mutant HPC2 target protein. Cells in positive wells are expanded and subcloned to establish and confirm monoclonality.

Clones with the desired specificities are expanded and grown as ascites in mice or in a hollow fiber system to produce sufficient quantities of antibody for characterization and assay development.

EXAMPLE 18

Isolation of HPC2 Binding Peptides

Peptides that bind to the HPC2 gene product are isolated from both chemical and phage-displayed random peptide libraries as follows.

Fragments of the HPC2 gene product are expressed as GST and His-tag fusion proteins in both *E. coli* and SF9 cells. The fusion protein is isolated using either a glutathione matrix (for GST fusions proteins) or nickel chelation matrix (for His-tag fusion proteins). This target fusion protein preparation is either screened directly as described below, or eluted with glutathione or imidizole. The target protein is immobilized to either a surface such as polystyrene; or a resin such as agarose; or solid supports using either direct absorption, covalent linkage reagents such as glutaraldehyde, or linkage agents such as biotin-avidin.

Two types of random peptide libraries of varying lengths are generated: synthetic peptide libraries that may contain derivatized residues, for example by phosphorylation or myristylation, and phage-displayed peptide libraries which may be phosphorylated. These libraries are incubated with immobilized HPC1 gene product in a variety of physiological buffers. Next, unbound peptides are removed by repeated washes, and bound peptides recovered by a variety of elution reagents such as low or high pH, strong denaturants, glutathione, or imidizole. Recovered synthetic peptide mixtures are sent to commercial services for peptide micro-sequencing to identify enriched residues. Recovered phage are amplified, rescreened, plaque purified, and then sequenced to determined the identity of the displayed peptides.

Use of HPC1 binding peptides. Peptides identified from the above screens are synthesized in larger quantities as biotin conjugates by commercial services. These peptides are used in both solid and solution phase competition assays with HPC1 and its interacting partners

identified in yeast 2-hybrid screens. Versions of these peptides that are fused to membrane-permeable motifs (Lin et al., 1995; Rojas et al., 1996) will be chemically synthesized, added to cultured cells and the effects on growth, apoptosis, differentiation, cofactor response, and internal changes will be assayed.

EXAMPLE 19

Sandwich Assay for HPC2

Monoclonal antibody is attached to a solid surface such as a plate, tube, bead, or particle. Preferably, the antibody is attached to the well surface of a 96-well ELISA plate. 100 µl sample (e.g., serum, urine, tissue cytosol) containing the HPC2 peptide/protein (wild-type or mutant) is added to the solid phase antibody. The sample is incubated for 2 hrs at room temperature. Next the sample fluid is decanted, and the solid phase is washed with buffer to remove unbound material. 100 µL of a second monoclonal antibody (to a different determinant on the HPC2 peptide/protein) is added to the solid phase. This antibody is labeled with a detector molecule (e.g., 125-I, enzyme, fluorophore, or a chromophore) and the solid phase with the second antibody is incubated for two hrs at room temperature. The second antibody is decanted and the solid phase is washed with buffer to remove unbound material.

The amount of bound label, which is proportional to the amount of HPC2 peptide/protein present in the sample, is quantified. Separate assays are performed using monoclonal antibodies which are specific for the wild-type HPC2 as well as monoclonal antibodies specific for each of the mutations identified in HPC2.

While the invention has been disclosed in this patent application by reference to the details of preferred embodiments of the invention, it is to be understood that the disclosure is intended in an illustrative rather than in a limiting sense, as it is contemplated that modifications will readily occur to those skilled in the art, within the spirit of the invention and the scope of the appended claims.

LIST OF REFERENCES

Altschul SF, *et al.* (1990). *J. Mol. Biol.* **215**: 403-410.

Altschul SF, *et al.* (1997). *Nucl. Acids Res.* **25**:3389-3402.

Anand R (1992). Techniques for the Analysis of Complex Genomes, (Academic Press).

Anderson WF, *et al.* (1980). *Proc. Natl. Acad. Sci. USA* **77**:5399-5403.

Antoniou AC, *et al.* (2000). *Genet. Epidemiol.* **18**:173-190.

Armitage P (1955). *Biometrics* **11**:375-386.

Ausubel FM, *et al.* (1992). Current Protocols in Molecular Biology, (J. Wiley and Sons, NY).

Bandyopadhyay PK and Temin HM (1984). *Mol. Cell. Biol.* **4**:749-754.

Barbeyron T, *et al.* (1995). *Microbiology* **141**:2897-2904.

Bartel PL, *et al.* (1993). "Using the 2-hybrid system to detect protein-protein interactions." *In:* Cellular Interactions in Development: A Practical Approach, Oxford University Press, pp. 153-179.

Beaucage SL and Caruthers MH (1981). *Tetra. Letts.* **22**:1859-1862.

Berglund P, *et al.* (1993). *Biotechnology* **11**:916-920.

Berkner KL (1992). *Curr. Top. Microbiol. Immunol.* **158**:39-66.

Berkner KL, *et al.* (1988). *BioTechniques* **6**:616-629.

Berry R, *et al.* (2000). *Am. J. Hum. Genet.* **66**:539-546.

Berthon P, *et al.* (1998). *Am. J. Hum. Genet.* **62**:1416-1424.

Borman S (1996). *Chemical & Engineering News*, December 9 issue, pp. 42-43.

Bouchardy C, *et al.* (1998). *Pharmacogenetics* **8**:291-298.

Bratt O, *et al.* (1999). *Br. J. Cancer* **81**:672-676.

Breakefield XO and Geller AI (1987). *Mol. Neurobiol.* **1**:337-371.

Breast Cancer Linkage Consortium (1999). *J. Natl. Cancer Inst.* **91**:1310-1316.

Brinster RL, *et al.* (1981). *Cell* **27**:223-231.

Buchschacher GL and Panganiban AT (1992). *J. Virol.* **66**:2731-2739.

Cannon L, *et al.* (1982). *Cancer Surveys* **1**:47-69.

Capecchi MR (1989). *Science* **244**:1288-1292.

Carfi A, *et al.* (1998). *Acta Crystallogr. D Biol. Crystallogr.* **54**:45-57.

Cariello NF (1988). *Am. J. Human Genetics* **42**:726-734.

Carter BS, *et al.* (1992). *Proc. Natl. Acad. Sci. USA* 89:3367-3371.

Carter BS, *et al.* (1993). *J. Urol.* **150**:797-802.

Chamberlain NL, *et al.* (1994). *Nucl. Acids Res.* **22**:3181-3186.

Chanfreau G, *et al.* (1996). *Science* **274**:1511-1514.

Chee M, *et al.* (1996). *Science* **274**:610-614.

Chevray PM and Nathans DN (1992). *Proc. Natl. Acad. Sci. USA* **89**:5789-5793.

Cochran WG (1954). *Biometrics* **10**:417-451.

Compton J (1991). *Nature* **350**:91-92.

Conner BJ, *et al.* (1983). *Proc. Natl. Acad. Sci. USA* **80**:278-282.

Cooney KA, *et al.* (1997). *J. Natl. Cancer Inst.* **89**:955-959.

Costantini F and Lacy E (1981). *Nature* **294**:92-94.

Cotten M, *et al.* (1990). *Proc. Natl. Acad. Sci. USA* **87**:4033-4037.

Cottingham RW, *et al.* (1993). *Am. J. Hum. Genet.* **53**:252-263.

Cotton RG, *et al.* (1988). *Proc. Natl. Acad. Sci. USA* **85**:4397-4401.

Couch FJ, *et al.* (1996). *Genomics* **36**:86-99.

Culver KW, *et al.* (1992). *Science* **256**:1550-1552.

Curiel DT, *et al.* (1991). *Proc. Natl. Acad. Sci. USA* **88**:8850-8854.

Curiel DT, *et al.* (1992). *Hum. Gene Ther.* **3**:147-154.

DeRisi J, *et al.* (1996). *Nature Genetics* **14**:457-460.

Deutscher, M (1990). *Meth. Enzymology* **182**:83-89 (Academic Press, San Diego, Cal.).

Donehower LA, *et al.* (1992). *Nature* **356**:215-221.

Durbin R and Thierry-Mieg J (1991). A *C. elegans* Database. Documentation, code and data available from anonymous FTP servers at lirmm.lirmm.fr, cele.mrc-lmb.cam.ac.uk and ncbi.nlm.nih.gov.

Editorial (1996). *Nature Genetics* **14**:367-370.

Eeles RA, *et al.* (1998). *Am. J. Hum. Genet.* **62**:653-658.

Elghanian R, *et al.* (1997). *Science* **277**:1078-1081.

Enhancers and Eukaryotic Gene Expression, Cold Spring Harbor Press, Cold Spring Harbor, New York (1983).

Erickson J, *et al.* (1990). *Science* **249**:527-533.

Fabiane SM, *et al.* (1998). *Biochemistry* **37**:12404-12411.

Fahy E, *et al.* (1991). *PCR Methods Appl.* **1**:25-33.

Feil R, *et al.*, (1996). *Proc. Natl. Acad. Sci. USA* **93**:10887-10890.

Felgner PL, *et al.* (1987). *Proc. Natl. Acad. Sci. USA* **84**:7413-7417.

Fields S and Song O-K (1989). *Nature* **340**:245-246.

Fiers W, *et al.* (1978). *Nature* **273**:113-120.

Fincham SM, *et al.* (1990). *The Prostate* **17**:189-206.

Fink DJ, *et al.* (1992). *Hum. Gene Ther.* **3**:11-19.

Fink DJ, *et al.* (1996). *Ann. Rev. Neurosci.* **19**:265-287.

Finkelstein J, *et al.* (1990). *Genomics* **7**:167-172.

Fodor SPA (1997). *Science* **277**:393-395.

Ford D, *et al.* (1998). *Am. J. Hum. Genet.* **62**:676-689.

Freese A, *et al.* (1990). *Biochem. Pharmacol.* **40**:2189-2199.

Friedman T (1991). In: Therapy for Genetic Diseases, T. Friedman, ed., Oxford University Press, pp. 105-121.

Gagneten S, *et al.* (1997). *Nucl. Acids Res.* **25**:3326-3331.

Gibbs M, *et al.* (1999a). *Am. J. Hum. Genet.* **64**:776-787.

Gibbs M, *et al.* (1999b). *Am. J. Hum. Genet.* **64**:1087-1095.

Giovannucci E, *et al.* (1997). *Proc. Natl. Acad. Sci. USA* 94:3320-3323.

Glover D (1985). DNA Cloning, **I** and **II** (Oxford Press).

Goding (1986). Monoclonal Antibodies: Principles and Practice, 2d ed. (Academic Press, NY).

Godowski PJ, *et al.* (1988). *Science* **241**:812-816.

Goldgar DE, *et al.* (1994). *J. Natl. Can. Inst.* **86**:3:200-209.

Goode EL, *et al.* (2000). *Genet. Epidemiol.* **18**:251-275.

Gordon JW, *et al.* (1980). *Proc. Natl. Acad. Sci. USA* **77**:7380-7384.

Gordon JW (1989). *Intl. Rev. Cytol.* **115**:171-229.

Gorziglia M and Kapikian AZ (1992). *J. Virol.* **66**:4407-4412.

Graham FL and van der Eb AJ (1973). *Virology* **52**:456-467.

Grompe M (1993). *Nature Genetics* **5**:111-117.

Grompe M, *et al.* (1989). *Proc. Natl. Acad. Sci. USA* **86**:5855-5892.

Gu H, *et al.* (1994). *Science* **265**:103-106.

Guthrie G and Fink GR (1991). Guide to Yeast Genetics and Molecular Biology (Academic Press).

Haase E, *et al.* (1989). *Mol. Gen. Genet.* **218**:64-71.

Hacia JG, *et al.* (1996). *Nature Genetics* **14**:441-447.

Hall JM, *et al.* (1990). *Science* **250**:1684-1689.

Harlow E and Lane D (1988). Antibodies: A Laboratory Manual (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY).

Harty LC, *et al.* (1997). *J. Natl. Cancer Inst.* **89**:1698-1705.

Hasty P, *et al.* (1991). *Nature* **350**:243-246.

Helseth E, *et al.* (1990). *J. Virol.* **64**:2416-2420.

Hodgson J (1991). *Bio/Technology* **9**:19-21.

Hori H, *et al.* (1997). *J. Clin. Gastroenterol.* **25**:568-575.

Hubert A, *et al.* (1999). *Am. J. Hum. Genet.* **65**:921-924.

Huse WD, *et al.* (1989). *Science* **246**:1275-1281.

Innis MA, *et al.* (1990). PCR Protocols: A Guide to Methods and Applications (Academic Press, San Diego, CA).

Jablonski E, *et al.* (1986). *Nucl. Acids Res.* **14**:6115-6128.

Jaffe JM, *et al.* (2000). *Cancer Res.* **60**:1626-1630.

Jakoby WB and Pastan IH (eds.) (1979). Cell Culture. *Methods in Enzymology*, Vol. 58 (Academic Press, Inc., Harcourt Brace Jovanovich (NY)).

Jenny A, *et al.* (1994). *Mol. Cell. Biol.* **14**:8183-8190.

Jenny A, *et al.* (1996). *Science* **274**:1514-1517.

Johnson PA, *et al.* (1992). *J. Virol.* **66**:2952-2965.

Johnson, *et al.* (1993). "Peptide Turn Mimetics" In: Biotechnology and Pharmacy, Pezzuto et al., eds., Chapman and Hall, NY.

Kaneda Y, *et al.* (1989). *J. Biol. Chem.* **264**:12126-12129.

Kanehisa M (1984). *Nucl. Acids Res.* **12**:203-213.

Kazemi-Esfarjani P, *et al.* (1995). *Hum. Mol. Genet.* **4**:523-527.

Kinszler KW, *et al.* (1991). *Science* **251**:1366-1370.

Kohler G and Milstein C (1975). *Nature* **256**:495-497.

Krain LS (1974). *Preventive Medicine* **3**:154-159.

Kubo T, *et al.* (1988). *FEBS Lett.* **241**:119-125.

Kyte J and Doolittle RF (1982). *J. Mol. Biol.* **157**:105-132.

Landegren U, *et al.* (1988). *Science* **242**:229-237.

Lander ES and Green P (1987). *Proc. Natl. Acad. Sci. USA* **84**:2363-2367.

Lange EM, *et al.* (1999). *Clin. Cancer Res.* **5**:4013-4020.

Lasko M, *et al.* (1992). *Proc. Natl. Acad. Sci. USA* 89:6232-6236.

Lathrop GM (1984). *Proc. Natl. Acad. Sci. USA* **81**:3443-3446.

Lavitrano M, *et al.* (1989). *Cell* **57**:717-723.

Lee JE, *et al.* (1995). *Science* **268**:836-844.

Lim CS, *et al.* (1991). *Circulation* **83**:2007-2011.

Lin YZ, *et al.* (1995). *J. Biol. Chem.* **270**:14255-14258.

Lipshutz RJ, *et al.* (1995). *BioTechniques* **19**:442-447.

Lo CW (1983). *Mol. Cell. Biol.* **3**:1803-1814.

Lobe CG and Nagy A (1998). *Bioessays* **20**:200-208.

Lockhart DJ, *et al.* (1996). *Nature Biotechnology* **14**:1675-1680.

Madzak C, *et al.* (1992). *J. Gen. Virol.* **73**:1533-1536.

Makridakis N, *et al.* (1997). *Cancer Res.* **57**:1020-1022.

Makridakis NM, *et al.* (1999). *Lancet* **354**:975-978.

Maniatis T, *et al.* (1982). Molecular Cloning: A Laboratory Manual (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY).

Mann R and Baltimore D (1985). *J. Virol.* **54**:401-407.

Margolskee RF (1992). *Curr. Top. Microbiol. Immunol.* **158**:67-95.

Martin R, *et al.* (1990). *BioTechniques* **9**:762-768.

Matteucci MD and Caruthers MH (1981). *J. Am. Chem. Soc.* **103**:3185.

Matthews JA and Kricka LJ (1988). *Anal. Biochem.* **169**:1.

Meikle AW, *et al.* (1985). *Prostate* **6**:121-128.

Melino S, *et al.* (1998). *TIBS* **23**:381-382.

Meniel V, *et al.* (1995). *Mutagenesis* **10**:543-548.

Merrifield B (1963). *J. Am. Chem. Soc.* **85**:2149-2156.

Metzger D, *et al.* (1988). *Nature* **334**:31-36.

Mifflin TE (1989). *Clinical Chem.* **35**:1819-1825.

Miki Y, *et al.* (1994). *Science* **266**:66-71.

Miller AD (1992). *Curr. Top. Microbiol. Immunol.* **158**:1-24.

Miller AD, *et al.* (1985). *Mol. Cell. Biol.* **5**:431-437.

Miller AD, *et al.* (1988). *J. Virol.* **62**:4337-4345.

Modrich P (1991). *Ann. Rev. Genet.* **25**:229-253.

Mombaerts P, *et al.* (1992). *Cell* **68**:869-877.

Morganti G, *et al.* (1956). *Acta Geneticae Medicae et Gemellogogiae* **6**:304-305.

Moss B (1992). *Curr. Top. Microbiol. Immunol.* **158**:25-38.

Moss B (1996). *Proc. Natl. Acad. Sci. USA* **93**:11341-11348.

Muzyczka N (1992). *Curr. Top. Microbiol. Immunol.* **158**:97-129.

Nabel (1992). *Hum. Gene Ther.* **3**:399-410.

Nabel EG, *et al.* (1990). *Science* **249**:1285-1288.

Naldini L, *et al.* (1996). *Science* **272**:263-267.

Nastiuk KL, *et al.* (1999). *Prostate* **40**:172-177.

Nevill-Manning CG, *et al.* (1998). *Proc. Natl. Acad. Sci. USA* **95**:5865-5871.

Neuhausen SL, *et al.* (1999). *Hum. Mol. Genet.* **8**:2437-2442.

Newton CR, *et al.* (1989). *Nucl. Acids Res.* **17**:2503-2516.

Nguyen Q, *et al.* (1992). *BioTechniques* **13**:116-123.

Niegemann E and Brendel M (1994). *Mutat. Res.* **315**:275-279.

Novack DF, *et al.* (1986). *Proc. Natl. Acad. Sci. USA* **83**:586-590.

O'Connell JR and Weeks DE (1995). *Nat. Genet.* **11**:402-408.

Ohi S, *et al.* (1990). *Gene* **89**:279-282.

Orita M, *et al.* (1989). *Proc. Natl. Acad. Sci. USA* **86**:2776-2770.

Osterrieder N and Wolf E (1998). *Rev. Sci. Tech.* **17**:351-364.

Ott J (1986). *Genet. Epidemiol. Suppl.* **1**:251-257.

Page KA, *et al.* (1990). *J. Virol.* **64**:5270-5276.

Page RDM (1996). *Computer Applications in the Biosciences* **12**:357-358.

Pellicer A, *et al.* (1980). *Science* **209**:1414-1422.

Peto J, *et al.* (1999). *J. Natl. Cancer Inst.* **91**:943-949.

Petropoulos CJ, *et al.* (1992). *J. Virol.* **66**:3391-3397.

Philpott KL, *et al.* (1992). *Science* **256**:1448-1452.

Quantin B, *et al.* (1992). *Proc. Natl. Acad. Sci. USA* **89**:2581-2584.

Remington's Pharmaceutical Sciences, 18th Ed. (1990, Mack Publishing Co., Easton, PA).

Rigby PWJ, *et al.* (1977). *J. Mol. Biol.* **113**:237-251.

Rojas M, *et al.* (1996). *J. Biol. Chem.* **271**:27456-27461.

Rosenfeld MA, *et al.* (1992). *Cell* **68**:143-155.

Ruano G and Kidd KK (1989). *Nucl. Acids Res.* **17**:8392.

Russell D and Hirata R (1998). *Nature Genetics* **18**:323-328.

Saitou N and Nei M (1987). *Mol. Biol. Evol.* **4**:406-425.

Sambrook J, *et al.* (1989). Molecular Cloning: A Laboratory Manual, 2nd Ed. (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY).

Schaffer AA, *et al.* (1994). *Hum. Hered.* **44**:225-237.

Scharf SJ (1986). *Science* **233**:1076-1078.

Schneider G, *et al.* (1998). *Nature Genetics* **18**:180-183.

Scopes R (1982). Protein Purification: Principles and Practice, (Springer-Verlag, NY).

Shastry BS (1995). *Experientia* **51**:1028-1039.

Shastry BS (1998). *Mol. Cell. Biochem.* **181**:163-179.

Sheffield VC, *et al.* (1989). *Proc. Natl. Acad. Sci. USA* **86**:232-236.

Sheffield VC, *et al.* (1991). *Am. J. Hum. Genet.* **49**:699-706.

Shenk TE, *et al.* (1975). *Proc. Natl. Acad. Sci. USA* **72**:989-993.

Shields PB (1997). *Proc. Dept. Defense BCRP Era of Hope* meeting, Vol. 1 ("Frontiers in Prevention and Detection"), pp.9-10.

Shimada T, *et al.* (1991). *J. Clin. Invest.* **88**:1043-1047.

Shinkai Y, *et al.* (1992). *Cell* **68**:855-867.

Shoemaker DD, *et al.* (1996). *Nature Genetics* **14**:450-456.

Sigurdsson S, *et al.* (1997). *J. Mol. Med.* **75**:758-761.

Smith JR, *et al.* (1996). *Science* **274**:1371-1374.

Smith SW, *et al.* (1994). *CABIOS* **10**:671-675.

Snouwaert JN, *et al.* (1992). *Science* **257**:1083-1088.

Sorge J, *et al.* (1984). *Mol. Cell. Biol.* **4**:1730-1737.

Spargo CA, *et al.* (1996). *Mol. Cell. Probes* **10**:247-256.

Stanford JL, *et al.* (1997). *Cancer Res.* **57**:1194-1198.

Steinberg GD, *et al.* (1990). *Prostate* **17**:337-347.

Stewart MJ, *et al.* (1992). *Hum. Gene Ther.* **3**:267-275.

Stratford-Perricaudet LD, *et al.* (1990). *Hum. Gene Ther.* **1**:241-256.

Suarez BK, *et al.* (2000). *Am. J. Hum. Genet.* **66**:933-944.

Tatusov RL, *et al.* (1997). *Science* **278**:631-637.

Tavtigian SV, *et al.* (1996). *Nat. Genet.* **12**:333-337.

Thierry-Mieg D, *et al.* (1995). Ace.mbly. A graphic interactive program to support shotgun and directed sequencing projects.

Thomas A, *et al.* (2000). *Statistics and Computing* **In press**.

Thompson JD, *et al.* (1997). *Nucl. Acids Res.* **25**:4876-4882.

Thompson S, *et al.* (1989). *Cell* **56**:313-321.

Valancius V and Smithies O (1991). *Mol. Cell Biol.* **11**:1402-1408.

Van der Putten H, *et al.* (1985). *Proc. Natl. Acad. Sci. USA* **82**:6148-6152.

Wagner E, *et al.* (1990). *Proc. Natl. Acad. Sci. USA* **87**:3410-3414.

Wagner E, *et al.* (1991). *Proc. Natl. Acad. Sci. USA* **88**:4255-4259.

Walker GT, *et al.* (1992). *Nucl. Acids Res.* **20**:1691-1696.

Wang CY and Huang L (1989). *Biochemistry* **28**:9508-9514.

Wartell RM, *et al.* (1990). *Nucl. Acids Res.* **18**:2699-2705.

Wells JA (1991). *Methods in Enzymol.* **202**:390-411.

Wetmur JG and Davidson N (1968). *J. Mol. Biol.* **31**:349-370.

White MB, *et al.* (1992). *Genomics* **12**:301-306.

White R and Lalouel JM (1988). *Annu. Rev. Genet.* **22**:259-279.

Wilkens EP, *et al.* (1999). *Prostate* **39**:280-284.

Wilkinson GW and Akrigg A (1992). *Nucleic Acids Res.* **20**:2233-2239.

Wolff JA, *et al.* (1990). *Science* **247**:1465-1468.

Wolff JA, *et al.* (1991). *BioTechniques* **11**:474-485.

Woolf CM (1960a). *Cancer* **13**:361-364.

Woolf CM (1960b). *Cancer* **13**:739-744.

Wooster R, *et al.* (1994). *Science* **265**:2088-2090.

Wooster R, *et al.* (1995). *Nature* **378**:789-792.

Wu DY and Wallace RB (1989). *Genomics* **4**:560-569.

Wu CH, *et al.* (1989). *J. Biol. Chem.* **264**:16985-16987.

Wu GY, *et al.* (1991). *J. Biol. Chem.* **266**:14338-14342.

Xu J, *et al.* (1998). *Nat. Genet.* **20**:175-179.

Xu J (2000). *Am. J. Hum. Genet.* 66:945-957.

Zenke M, *et al.* (1990). *Proc. Natl. Acad. Sci. USA* **87**:3655-3659.

U.S. Patent No. 3,817,837

U.S. Patent No. 3,850,752

U.S. Patent No. 3,939,350

U.S. Patent No. 3,996,345

U.S. Patent No. 4,275,149

U.S. Patent No. 4,277,437

U.S. Patent No. 4,366,241

U.S. Patent No. 4,376,110

U.S. Patent No. 4,486,530

U.S. Patent No. 4,554,101

U.S. Patent No. 4,683,195

U.S. Patent No. 4,683,202

U.S. Patent No. 4,816,567

U.S. Patent No. 4,868,105

U.S. Patent No. 4,873,191

U.S. Patent No. 5,252,479

U.S. Patent No. 5,270,184

U.S. Patent No. 5,409,818

U.S. Patent No. 5,436,146

U.S. Patent No. 5,455,166

U.S. Patent No. 5,550,050

U.S. Patent No. 5,691,198

U.S. Patent No. 5,735,500

U.S. Patent No. 5,747,469

Hitzeman *et al.*, EP 73,675A

EPO Publication No. 225,807

EP 425,731A

European Patent Application Publication No. 0332435

WO 84/03564

WO 90/07936
WO 92/19195
WO 93/07282
WO 94/25503
WO 95/01203
WO 95/05452
WO 96/02286
WO 96/02646
WO 96/11698
WO 96/40871
WO 96/40959
WO 97/12635